Cellular Tree Classifiers

Gérard Biau, Luc Devroye

▶ **To cite this version:**

Gérard Biau, Luc Devroye. Cellular Tree Classifiers. 2013. hal-00778520v2

HAL Id: hal-00778520
https://hal.science/hal-00778520v2

Preprint submitted on 24 Jun 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Cellular Tree Classifiers

Gérard Biau
Université Pierre et Marie Curie[1] *& Ecole Normale Supérieure*[2]*, France*
gerard.biau@upmc.fr

Luc Devroye
McGill University, Canada[3]
lucdevroye@gmail.com

Abstract

The cellular tree classifier model addresses a fundamental problem in the design of classifiers for a parallel or distributed computing world: Given a data set, is it sufficient to apply a majority rule for classification, or shall one split the data into two or more parts and send each part to a potentially different computer (or cell) for further processing? At first sight, it seems impossible to define with this paradigm a consistent classifier as no cell knows the "original data size", n. However, we show that this is not so by exhibiting two different consistent classifiers. The consistency is universal but is only shown for distributions with nonatomic marginals.

Index Terms — Classification, pattern recognition, tree classifiers, cellular computation, Bayes risk consistency, asymptotic analysis, nonparametric estimation.

2010 Mathematics Subject Classification: 62G05, 62G20.

1 Introduction

1.1 The problem

We explore in this paper a new way of dealing with the supervised classification problem. In the model we have in mind, a basic computational unit in classification, a cell, takes as input training data, and makes a decision whether a majority rule should be applied to all data, or whether the data should be split, and each part of the partition should be given to another cell.

[1]Research partially supported by the French National Research Agency (grant ANR-09-BLAN-0051-02 "CLARA") and by the Institut universitaire de France.

[2]Research carried out within the INRIA project "CLASSIC" hosted by Ecole Normale Supérieure and CNRS.

[3]Research sponsored by NSERC Grant A3456 and FQRNT Grant 90-ER-0291.

All cells must be the same—their function is not altered by external inputs. In other words, the decision to split depends only upon the data presented to the cell. Classifiers designed according to this autonomous principle will be called cellular tree classifiers, or simply cellular classifiers. This manner of tackling the classification problem is novel, but has a wide reach in a world in which parallel and distributed computation are important. In the short term, parallelism will take hold in massive data sets and complex systems and, as such, is one of the exciting questions that will be asked to the statistics and machine learning fields.

The purpose of the present document is to formalize the setting and to provide a foundational discussion of various properties, good and bad, of tree classifiers that are formulated following these principles. Our constructions lead to classifiers that always converge. They are the first consistent cellular classifiers that we are aware of. This article is also motivated by the challenges involved in "big data" issues (see, e.g., Jordan, 2011), in which recursive approaches such as divide-and-conquer algorithms (e.g., Cormen et al., 2009) play a central role. Such procedures are naturally adapted for execution in multi-processor machines, especially shared-memory systems where the communication of data between processors does not need to be planned in advance.

In the design of classifiers, we have an unknown distribution of a random prototype pair $(\mathbf{X}, Y)$, where $\mathbf{X}$ takes values in $\mathbb{R}^d$ and Y takes only finitely many values, say 0 or 1 for simplicity. Classical pattern recognition deals with predicting the unknown nature Y of the observation $\mathbf{X}$ via a measurable classifier $g : \mathbb{R}^d \to \{0, 1\}$. Since it is not assumed that $\mathbf{X}$ fully determines the label, it is certainly possible to misspecify its associated class. Thus, we err if $g(\mathbf{X})$ differs from Y, and the probability of error for a particular decision rule g is $L(g) = \mathbb{P}\{g(\mathbf{X}) \neq Y\}$. The Bayes classifier

$$g^\star(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbb{P}\{Y = 1|\mathbf{X} = \mathbf{x}\} > \mathbb{P}\{Y = 0|\mathbf{X} = \mathbf{x}\} \\ 0 & \text{otherwise} \end{cases}$$

has the smallest probability of error, that is

$$L^\star = L(g^\star) = \inf_{g:\mathbb{R}^d \to \{0,1\}} \mathbb{P}\{g(\mathbf{X}) \neq Y\}$$

(see, for instance, Theorem 2.1 in Devroye et al., 1996). However, most of the time, the distribution of $(\mathbf{X}, Y)$ is unknown, so that $g^\star$ is unknown too. Fortunately, it is often possible to collect a sample (the data) $\mathcal{D}_n = ((\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n))$ of independent and identically distributed (i.i.d.) copies of $(\mathbf{X}, Y)$. We assume that $\mathcal{D}_n$ and $(\mathbf{X}, Y)$ are independent. In this

context, a classifier $g_n(\mathbf{x}; \mathcal{D}_n)$ is a measurable function of $\mathbf{x}$ and $\mathcal{D}_n$, and it attempts to estimate Y from $\mathbf{X}$ and $\mathcal{D}_n$. For simplicity, we suppress $\mathcal{D}_n$ in the notation and write $g_n(\mathbf{x})$ instead of $g_n(\mathbf{x}; \mathcal{D}_n)$.

The probability of error of a given classifier g_n is the random variable

$$L(g_n) = \mathbb{P}\{g_n(\mathbf{X}) \neq Y | \mathcal{D}_n\},$$

and the rule is consistent if

$$\lim_{n \to \infty} \mathbb{E}L(g_n) = L^\star.$$

It is universally consistent if it is consistent for all possible distributions of $(\mathbf{X}, Y)$. Many popular classifiers are universally consistent. These include several brands of histogram rules, k-nearest neighbor rules, kernel rules, neural networks, and tree classifiers. There are too many references to be cited here, but the monographs by Devroye et al. (1996) and Györfi et al. (2002) will provide the reader with a comprehensive introduction to the domain and a literature review. Among these rules, tree methods loom large for several reasons. All procedures that partition space, such as histogram rules, can be viewed as special cases of partitions generated by trees. Simple neural networks that use voting methods can also be regarded as trees, and similarly, kernel methods with kernels that are indicator functions of sets are but special cases of tree methods. Tree classifiers are conceptually simple, and explain the data very well. However, their design can be cumbersome, as optimizations performed over all possible tree classifiers that follow certain restrictions could face a huge combinatorial and computational hurdle. The cellular paradigm addresses these concerns.

Partitions of $\mathbb{R}^d$ based upon trees have been studied in the computational geometry literature (Bentley, 1975; Overmars and van Leeuwen, 1982; Edelsbrunner and van Leeuwen, 1983; Mehlhorn, 1984) and the computer graphics literature (Samet, 1984, 1990). Most popular among these are the k-d trees and quadtrees. Our version of space partitioning corresponds to Bentley's k-d trees (1975). The basic notions of trees as related to pattern recognition can be found in Chapter 20 of Devroye et al. (1996). However, trees have been suggested as tools for classification more than twenty years before that. We mention in particular the early work of Fu (You and Fu, 1976; Anderson and Fu, 1979; Mui and Fu, 1980; Lin and Fu, 1983; Qing-Yun and Fu, 1983). Other references from the 1970s include Meisel and Michalopoulos (1973); Bartolucci et al. (1976); Payne and Meisel (1977); Sethi and Chatterjee (1977); Swain and Hauska (1977); Gordon and Olshen (1978); Friedman (1979). Most influential in the classification tree literature was the CART

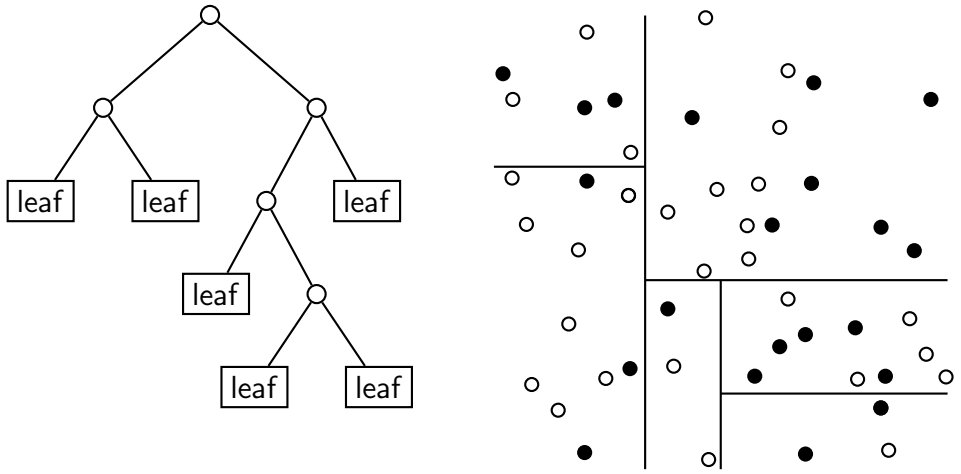

Figure 1: A binary tree (left) and the corresponding partition (right).

proposal by Breiman et al. (1984). While CART proposes partitions by hyperrectangles, linear hyperplanes in general position have also gained in popularity—the early work on that topic is by Loh and Vanichsetakul (1988), and Park and Sklansky (1990). Additional references on tree classification include Gustafson et al. (1980); Argentiero et al. (1982); Hartmann et al. (1982); Kurzynski (1983); Wang and Suen (1984); Suen and Wang (1987); Shlien (1990); Chou (1991); Gelfand and Delp (1991); Gelfand et al. (1991); Simon (1991); Guo and Gelfand (1992).

1.2 The cellular computation spirit

In general, classification trees partition $\mathbb{R}^d$ into regions, often hyperrectangles parallel to the axes (an example is depicted in Figure 1). In t-ary trees, each node has exactly t or 0 children. If a node u represents the set A and its children $u_1, \ldots, u_t$ represent $A_1, \ldots, A_t$, then it is required that $A = A_1 \cup \cdots \cup A_t$ and $A_i \cap A_j = \emptyset$ for $i \neq j$. The root of the tree represents $\mathbb{R}^d$, and the terminal nodes (or leaves), taken together, form a partition of $\mathbb{R}^d$. If a leaf represents region A, then the tree classifier takes the simple form

$$g_n(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i=1}^n \mathbf{1}_{[\mathbf{X}_i \in A, Y_i = 1]} > \sum_{i=1}^n \mathbf{1}_{[\mathbf{X}_i \in A, Y_i = 0]}, \quad \mathbf{x} \in A \\ 0 & \text{otherwise.} \end{cases}$$

That is, in every leaf region, a majority vote is taken over all $(\mathbf{X}_i, Y_i)$'s with $\mathbf{X}_i$'s in the same region. Ties are broken, by convention, in favor of class 0.

The tree structure is usually data-dependent, as well, and indeed, it is in the construction itself where different trees differ. Thus, there are virtually

infinitely many possible strategies to build classification trees. Nevertheless, despite this great diversity, all tree species end up with two fundamental questions at each node:

> ① Should the node be split?
>
> ② In the affirmative, what are its children?

These two questions are typically answered using **global** information regarding the tree, such as, for example, a function of the data $\mathcal{D}_n$, the level of the node within the tree, the size of the data set and, more generally, any parameter connected with the structure of the tree. This parameter could be, for example, the total number k of cells in a k-partition tree or the penalty term in the pruning of the CART algorithm (Breiman et al., 1984; see also Gey and Nédélec, 2005).

Cellular trees proceed from a different philosophy. In short, a cellular tree should, at each node, be able to answer questions ① and ② using **local** information only, without any help from the other nodes. In other words, each cell can perform as many operations it wishes, provided it uses **only** the data that are transmitted to it, regardless of the general structure of the tree. Just imagine that the calculations to be carried out at the nodes are sent to different computers, eventually asynchronously, and that the system architecture is so complex that computers do not communicate. Such a situation may arise, for example, in the context of massive data sets, that is, when both n and d are astronomical, and no single human and no single computer can handle this alone. Thus, once a computer receives its data, it has to make its own decisions ① and ② based on this data subset only, independently of the others and without knowing anything of the overall edifice. Once a data set is split, it can be given to another computer for further splitting, since the remaining data points have no influence. This greedy mechanism is schematized in Figure 2.

But there is a more compelling reason for making local decisions. A neurologist seeing twenty patients must make decisions without knowing anything about the other patients in the hospital that were sent to other specialists. Neither does he need to know how many other patients there are. The neurologist's decision, in other words, should only be based on the data—the patients—in his care.

Decision tree learning is a method commonly used in data mining (see, e.g., Rokach and Maimon, 2008). Its goal is to create a model that partitions the

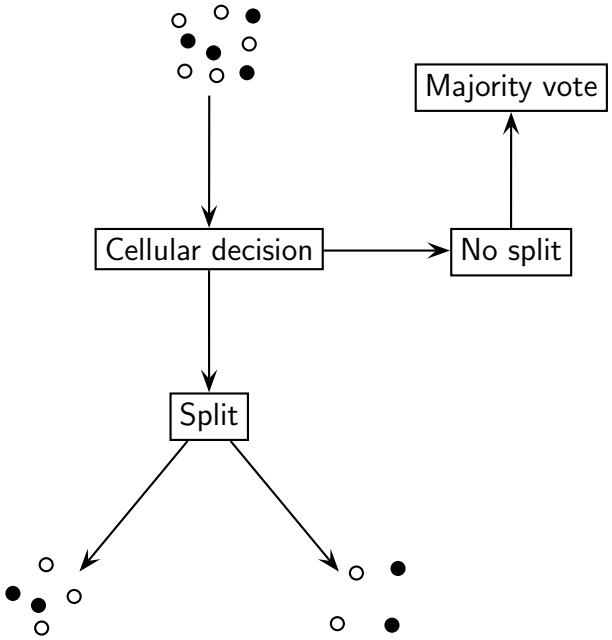

Figure 2: Schematization of the cell, the computational unit.

space recursively, as in a tree, in which leaf nodes (terminal nodes) correspond to final decisions. This process of top-down induction of decision trees—a phrase introduced by Quinlan in 1968—is called greedy in the data mining and computer science literature. It is by far the most common strategy for learning decision trees from data. The literature on this topic is largely concerned with the manner in which splits are made, and with the stopping rule.

For example, in CART (Breiman et al., 1984), splits are made perpendicular to the axes based on the notion of Gini impurity. Splits are performed until all data are isolated. In a second phase, nodes are recombined from the bottom-up in a process called pruning. It is this second process that makes the CART trees non-cellular, as global information is shared to manage the recombination process. Quinlan's C4.5 (1993) also prunes. Others split until all nodes or cells are homogeneous (i.e., have the same class)—the prime example is Quinlan's ID3 (1986). This strategy, while compliant with the cellular framework, leads to non-consistent rules, as we point out in the present paper. In fact, the choice of a good stopping rule for decision trees is very hard—we were not able to find any in the literature that guarantee convergence to the Bayes error.

We note here that decision networks have received renewed attention in wireless sensor networks (see, e.g., Arora et al., 2004, or Cheng et al., 2010). Phys-

ical and energy considerations impose a natural restriction on the classifiers—decisions must be taken locally. This corresponds, in spirit, to the cellular framework we are proposing. However, most sensor network decision trees use global criteria such as pruning that are based on a global method of deciding where to prune. The consistency question has not been addressed in these applications.

2 Cellular tree classifiers

2.1 A mathematical model

The objective of this subsection is to discuss a tentative mathematical model for cellular tree classifiers. Without loss of generality, we consider binary tree classifiers based on a class $\mathcal{C}$ of possible Borel subsets of $\mathbb{R}^d$ that can be used for splits. A typical example of such a class is the family of all hyperplanes, or the class of all hyperplanes that are perpendicular to one of the axes. Higher order polynomial splitting surfaces can be imagined as well.

The class is parametrized by a vector $\sigma \in \mathbb{R}^p$. There is a splitting function $f(\mathbf{x}, \sigma)$, $\mathbf{x} \in \mathbb{R}^d, \sigma \in \mathbb{R}^p$, such that $\mathbb{R}^d$ is partitioned into $A = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}, \sigma) \geq 0\}$ and $B = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}, \sigma) < 0\}$. Formally, a cellular split can be viewed as a family of measurable mappings σ from $(\mathbb{R}^d \times \{0, 1\})^n$ to $\mathbb{R}^p$ (for all $n \geq 1$). That is, for each possible input size n, we have a map. In addition, there is a family of measurable mappings θ from $(\mathbb{R}^d \times \{0, 1\})^n$ to $\{0, 1\}$ that indicate decisions: $\theta = 1$ indicates that a split should be applied, while $\theta = 0$ corresponds to a decision not to split. In that case, the cell acts as a leaf node in the tree. Note that θ and σ correspond to the decisions given in ① and ②.

A cellular binary classification tree is a machine that partitions the space recursively in the following manner. With each node we associate a subset of $\mathbb{R}^d$, starting with $\mathbb{R}^d$ for the root node. Let the data set be $\mathcal{D}_n$. If $\theta(\mathcal{D}_n) = 0$, the root cell is final, and the space is not split. Otherwise, $\mathbb{R}^d$ is split into

$$A = \left\{\mathbf{x} \in \mathbb{R}^d : f\left(\mathbf{x}, \sigma(\mathcal{D}_n)\right) \geq 0\right\} \quad \text{and} \quad B = \left\{\mathbf{x} \in \mathbb{R}^d : f\left(\mathbf{x}, \sigma(\mathcal{D}_n)\right) < 0\right\}.$$

The data $\mathcal{D}_n$ are partitioned into two groups—the first group contains all $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, for which $\mathbf{X}_i \in A$, and the second group all others. The groups are sent to child cells, and the process is repeated.

A priori, there is no reason why this tree should be finite. We will impose conditions later on that ensure that with probability 1, the tree is finite for all n and for all possible values of the data. For example, this could be

achieved by hyperplane splits perpendicular to the axes that are forced to visit (contain) one of the $\mathbf{X}_i$'s. By insisting that the data point selected on the boundary be "eaten", i.e., not sent down to the child nodes, one reduces the data set by one at each split, thereby ensuring the finiteness of the decision tree. We will employ such a (crude) method.

When $\mathbf{x} \in \mathbb{R}^d$ needs to be classified, we first determine the unique leaf set $A(\mathbf{x})$ to which $\mathbf{x}$ belongs, and then take votes among the $\{Y_i : \mathbf{X}_i \in A(\mathbf{x}), i = 1, \ldots, n\}$. Classification proceeds by a majority vote, with the majority deciding the estimate $g_n(\mathbf{x})$. In case of a tie, we set $g_n(\mathbf{x}) = 0$.

A cellular binary tree classifier is said to be randomized if each node in the tree has an independent copy of a uniform $[0, 1]$ random variable associated with it, and θ and σ are mappings that have one extra real-valued component in the input. For example, we could flip an unbiased coin at each node to decide whether $\theta = 0$ or $\theta = 1$.

Remark 2.1 *It is tempting to say that any classifier g_n is a cellular tree classifier with the following mechanism: Set $\theta = 1$ if we are at the root, and $\theta = 0$ elsewhere. The root node is split by the classifier into a set*

$$A = \{\mathbf{x} \in \mathbb{R}^d : g_n(\mathbf{x}) = 1\}$$

and its complement, and both child nodes are leaves. However, the decision to cut can only be a function of the input data, and not the node's position in the tree, and thus, this is not allowed.

2.2 Are there consistent cellular tree classifiers?

At first sight, it appears that there are no universally consistent cellular tree classifiers. Consider for example complete binary trees with k full levels, i.e., there are 2^k leaf regions. We can have consistency when k is allowed to depend upon n. An example is the median tree (Devroye et al., 1996, Section 20.3). When $d = 1$, split by finding the median element among the $\mathbf{X}_i$'s, so that the child sets have cardinality given by $\lfloor (n-1)/2 \rfloor$ and $\lceil (n-1)/2 \rceil$, where $\lfloor . \rfloor$ and $\lceil . \rceil$ are the floor and ceiling functions. The median itself does stay behind and is not sent down to the subtrees, with an appropriate convention for breaking cell boundaries as well as empty cells. Keep doing this for k rounds—in d dimensions, one can either rotate through the coordinates for median splitting, or randomize by selecting uniformly at random a coordinate to split orthogonally.

This rule is known to be consistent as soon as the marginal distributions of $\mathbf{X}$ are nonatomic, provided $k \to \infty$ and $k2^k/n \to 0$. However, this is not

a cellular tree classifier. While we can indeed specify σ, it is impossible to define θ because θ cannot be a function of the global value of n. In other words, if we were to apply median splitting and decide to split for a fixed k, then the leaf nodes would all correspond to a fix proportion of the data points. It is clear that the decisions in the leaves are off with a fair probability if we have, for example, Y independent of $\mathbf{X}$ and $\mathbb{P}\{Y = 1\} = 1/2$. Thus, we cannot create a cellular tree classifier in this manner.

In view of the preceding discussion, it seems paradoxical that there indeed exist universally consistent cellular tree classifiers. (We note here that we abuse the word "universal"—we will assume throughout, to keep the discussion at a manageable level, that the marginal distributions of $\mathbf{X}$ are nonatomic. But no other conditions on the joint distribution of $(\mathbf{X}, Y)$ are imposed.) Our first construction, which is presented in Section 3, follows the median tree principle and uses randomization. In a second construction (Section 4) we derandomize, and exploit the idea that each cell is allowed to explore its own subtrees, thereby anticipating the decisions of its children. For the sake of clarity, proofs of the most technical results are gathered in Section 5 and Section 6.

3 A randomized cellular tree classifier

From now on, to keep things simple, it is assumed that the marginal distributions of $\mathbf{X}$ are nonatomic. The cellular splitting method σ described in this section mimics the median tree classifier discussed above. We first choose a dimension to cut, uniformly at random from the d dimensions, as rotating through the dimensions by level number would violate the cellular condition. The selected dimension is then split at the data median, just as in the classical median tree. Repeating this for k levels of nodes leads to 2^k leaf regions. On any path of length k to one of the 2^k leaves, we have a deterministic sequence of cardinalities $n_0 = n(\text{root}), n_1, n_2, \ldots, n_k$. We always have $n_i/2 - 1 \leq n_{i+1} \leq n_i/2$. Thus, by induction, one easily shows that, for all i,

$$\frac{n}{2^i} - 2 \leq n_i \leq \frac{n}{2^i}.$$

In particular, each leaf has at least $\max(n/2^k - 2, 0)$ points and at most $n/2^k$.

Remark 3.1 *The problem of atoms in the coordinates can be dealt with separately, but still within the cellular framework. The particularity is that the threshold for splitting may now be at a position at which one or more data values occur. This leaves two sets that may differ in size by more than one.*

The atoms in the distribution of $\mathbf{X}$ *can never be separated, but that is as it should be. We leave it to the reader to adapt the subsequent arguments to the case of atomic distributions.*

The novelty is in the choice of the decision function. This function ignores the data altogether and uses a randomized decision that is based on the size of the input. More precisely, consider a nonincreasing function $\varphi : \mathbb{N} \to (0, 1]$ with $\varphi(0) = \varphi(1) = 1$. Cells correspond in a natural way to sets of $\mathbb{R}^d$. So, we can and will speak of a cell A, where $A \subset \mathbb{R}^d$. The number of data points in A is denoted by $N(A)$:

$$N(A) = \sum_{i=1}^{n} \mathbf{1}_{[\mathbf{X}_i \in A]}.$$

Then, if U is the uniform $[0, 1]$ random variable associated with the cell A and the input to the cell is $N(A)$, the stopping rule ① takes the form:

> ① Put $\theta = 0$ if
>
> $$U \leq \varphi\left(N(A)\right).$$

In this manner, we obtain a possibly infinite randomized binary tree classifier. Splitting occurs with probability $1 - \varphi(n)$ on inputs of size n. Note that no attempt is made to split empty sets or singleton sets. For consistency, we need to look at the random leaf region to which $\mathbf{X}$ belongs. This is roughly equivalent to studying the distance from that cell to the root of the tree.

In the sequel, the notation $u_n = \mathrm{o}(v_n)$ (respectively, $u_n = \omega(v_n)$ and $u_n = \mathrm{O}(v_n)$) means that $u_n/v_n \to 0$ (respectively, $v_n/u_n \to 0$ and $u_n \leq C v_n$ for some constant C) as $n \to \infty$. Many choices $\varphi(n) = \mathrm{o}(1)$, but not all, will do for us. The next lemma makes things more precise.

Lemma 3.1 *Let* $\beta \in (0, 1)$. *Define*

$$\varphi(n) = \begin{cases} 1 & \text{if } n < 3 \\ 1/\log^{\beta} n & \text{if } n \geq 3. \end{cases}$$

Let $K(\mathbf{X})$ *denote the random path distance between the cell of* $\mathbf{X}$ *and the root of the tree. Then*

$$\lim_{n \to \infty} \mathbb{P}\left\{K(\mathbf{X}) \geq k_n\right\} = \begin{cases} 0 & \text{if } k_n = \omega(\log^{\beta} n) \\ 1 & \text{if } k_n = \mathrm{o}(\log^{\beta} n). \end{cases}$$

Proof of Lemma 3.1 Let us recall that, at level k, each cell of the underlying median tree contains at least $\max(n/2^k - 2, 0)$ points and at most $n/2^k$. Since the function $\varphi(.)$ is nonincreasing, the first result follows from this:

$$\begin{aligned}\mathbb{P}\left\{K(\mathbf{X}) \geq k_n\right\} &\leq \prod_{i=0}^{k_n-1} \left(1 - \varphi\left(\lfloor n/2^i \rfloor\right)\right) \\ &\leq \exp\left(-\sum_{i=0}^{k_n-1} \varphi\left(\lfloor n/2^i \rfloor\right)\right) \\ &\leq \exp\left(-k_n \varphi(n)\right).\end{aligned}$$

The second statement follows from

$$\mathbb{P}\left\{K(\mathbf{X}) < k_n\right\} \leq \sum_{i=0}^{k_n-1} \varphi\left(\lceil n/2^i - 2 \rceil\right) \leq k_n \varphi\left(\lceil n/2^{k_n} \rceil\right),$$

valid for all n large enough since $n/2^{k_n} \to \infty$ as $n \to \infty$. ■

Lemma 3.1, combined with the median tree consistency result of Devroye et al. (1996), suffices to establish consistency of the randomized cellular tree classifier.

Theorem 3.1 *Let β be a real number in $(0, 1)$. Define*

$$\varphi(n) = \begin{cases} 1 & \text{if } n < 3 \\ 1/\log^\beta n & \text{if } n \geq 3. \end{cases}$$

Let g_n be the associated randomized cellular binary tree classifier. Assume that the marginal distributions of $\mathbf{X}$ are nonatomic. Then the classification rule g_n is consistent:

$$\lim_{n\to\infty} \mathbb{E}L(g_n) = L^\star \quad \text{as } n \to \infty.$$

Proof of Theorem 3.1 By $\mathrm{diam}(A)$ we mean the diameter of the cell A, i.e., the maximal distance between two points of A. We recall a general consistency theorem for partitioning classifiers whose cell design depends on the $\mathbf{X}_i$'s only (Devroye et al., 1996, Theorem 6.1). According to this theorem, such a classifier is consistent if both

1. $\mathrm{diam}(A(\mathbf{X})) \to 0$ in probability as $n \to \infty$, and
2. $N(A(\mathbf{X})) \to \infty$ in probability as $n \to \infty$,

where $A(\mathbf{X})$ is the cell of the random partition containing $\mathbf{X}$.

Condition 2. is proved in Lemma 3.1. Notice that

$$\begin{aligned} N\left(A(\mathbf{X})\right) &\geq \frac{n}{2^{K(\mathbf{X})}} - 2 \\ &\geq \mathbf{1}_{[K(\mathbf{X})<\log^{(\beta+1)/2} n]} \left(\frac{n}{2^{\log^{(\beta+1)/2} n}} - 2 \right) \\ &= \omega(1)\mathbf{1}_{[K(\mathbf{X})<\log^{(\beta+1)/2} n]}. \end{aligned}$$

Therefore, by Lemma 3.1, $N\left(A(\mathbf{X})\right) \to \infty$ in probability as $n \to \infty$.

To show that $\mathrm{diam}(A(\mathbf{X})) \to 0$ in probability, observe that on a path of length $K(\mathbf{X})$, the number of times the first dimension is cut is binomial $(K(\mathbf{X}), 1/d)$. This tends to infinity in probability. Following the proof of Theorem 20.2 in Devroye et al. (1996), the diameter of the cell of $\mathbf{X}$ tends to 0 in probability with n. Details are left to the reader. ■

Let us finally take care of the randomization. Can one do without randomization? The hint to the solution of that enigma is in the hypothesis that the data elements in $\mathcal{D}_n$ are i.i.d. The median classifier does not use the ordering in the data. Thus, one can use the randomness present in the permutation of the observations, e.g., the ℓ-th components of the $\mathbf{X}_i$'s can form $n!$ permutations if ties do not occur. This corresponds to $(1 + \mathrm{o}(1))n \log_2 n$ independent fair coin flips, which are at our disposal. Each decision to split requires on average at most 2 independent bits. The selection of a random direction to cut requires no more than $1 + \log_2 d$ independent bits. Since the total tree size is, with probability tending to 1, $\mathrm{O}(2^{\log^{\beta+\varepsilon} n})$ for any $\varepsilon > 0$, a fact that follows with a bit of work from summing the expected number of nodes at each level, the total number of bits required to carry out all computations is

$$\mathrm{O}\left((3 + \log_2 d) 2^{\log^{\beta+\varepsilon} n} \right),$$

which is orders of magnitude smaller than n provided that $\beta + \varepsilon < 1$. Thus, there is sufficient randomness at hand to do the job. How it is actually implemented is another matter, as there is some inevitable dependence between the data sets that correspond to cells and the data sets that correspond to their children. We will not worry about the finer details of this in the present paper.

Remark 3.2 *For more on random tree models and their analyses, see the texts of Drmota (2009), and Flajolet and Sedgewick (2008). Additional material on information-theory and bit complexity can be found in the monograph by Cover and Thomas (2006).*

Remark 3.3 *In the spirit of Breiman's random forests (2001), one could envisage to use a collection of randomized cellular tree classifiers and make final predictions by aggregating over the ensemble. Since each individual rule is consistent (by Theorem 3.1), then the same property is also true for the ensemble (see, e.g., Proposition 1 in Biau et al., 2008). Improvements are expected at the level of predictive accuracy and stability.*

4 A non-randomized cellular tree classifier

The cellular tree classifier that we consider in this section is more sophisticated and autonomous, in the sense that it does not rely on any randomization scheme. It partitions the data recursively as follows. With each node we associate a set of $\mathbb{R}^d$, starting with $\mathbb{R}^d$ for the root node. We first consider a full 2^d-ary tree (see Figure 3 for an illustration in dimension 2), with the cuts decided in the following manner. The dimensions are ordered once and for all from 1 to d. At the root, we find the median of (the projection of) the n data points in direction 1, then on each of the two subsets, we find the median in direction 2, then on each of the four subsets, we find the median in direction 3, and so forth. A split, contrary to our discussion thus far, is into 2^d parts, not two parts. This corresponds to Bentley's k-d tree (1975). Repeating this splitting for k levels of nodes leads to 2^{dk} leaf regions, each having at least $\max(n/2^{dk} - 2, 0)$ points and at most $n/2^{dk}$.

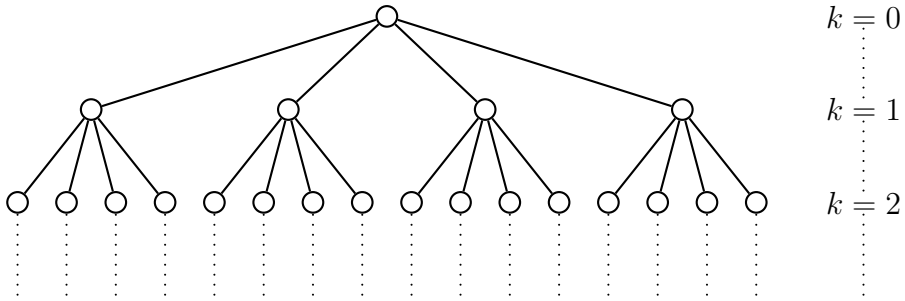

Figure 3: A full 2^d-ary tree in dimension $d = 2$.

This procedure is equivalent to dk consecutive binary splits at the median, where we rotate through the dimensions. However, in our cellular set-up, such rotations through the dimensions are impossible, and this forces us to employ this equivalent strategy. Note, therefore, that the split parameter σ is an extension of the binary classifier split σ—one could consider it as a vector

of dimension $2^d - 1$, as we need to specify $2^d - 1$ coordinate positions to fully specify a partition into 2^d regions. It remains to specify a stopping rule θ which respects the cellular constraint. To this aim, we need some additional notation.

Remark 4.1 *By the very construction of the tree, at each node, the median itself does stay behind and is not sent down to the subtrees. From a topological point of view, this means that, in the partition building, each cell A and its 2^d child cells $A_1, \ldots, A_{2^d}$ are considered as* **open** *hyperrectangles. Thus, for classification, assuming nonatomic marginals, we would thus strictly speaking not be able to classify any data that fall "on the border" between $A_1, \ldots, A_{2^d}$. This is a non-important detail for the calculations since the marginal distributions of $\mathbf{X}$ are nonatomic. In practice, this issue can be solved with an appropriate convention to break the boundary ties.*

If A is any cell of the full 2^d-ary tree defined above, we let $N(A)$ be the number of $\mathbf{X}_i$'s falling in A, and estimate the quality of the majority vote classifier at this node by

$$\hat{L}_n(A) = \frac{1}{N(A)} \min \left(\sum_{i=1}^{n} \mathbf{1}_{[\mathbf{X}_i \in A, Y_i = 1]}, \sum_{i=1}^{n} \mathbf{1}_{[\mathbf{X}_i \in A, Y_i = 0]} \right).$$

(Throughout, we adopt the convention $0/0 = 0$.)

Remark 4.2 *Each cut at the median eliminates 1 data point. Thus, given a cell A, the construction of its offspring k generations later rules out at most $1 + \cdots + 2^{dk-1} = 2^{dk} - 1$ observations. In particular, if A has cardinality $N(A)$, then, k generations later, its offspring $A_1, \ldots, A_{2^{dk}}$ have a total combined cardinality at least $N(A) - (2^{dk} + 1)$.*

Fix a positive real parameter α and define the nonnegative integer k^+ by

$$k^+ = \lfloor \alpha \log_2(N(A) + 1) \rfloor,$$

where, for simplicity, we drop the dependency of k^+ upon A and α. Finally, letting $\mathcal{P}_{k^+}(A)$ be the 2^{dk^+} leaf regions (terminal nodes) of the full 2^d-ary tree rooted at A of height k^+, we set

$$\hat{L}_n(A, k^+) = \sum_{A_j \in \mathcal{P}_{k^+}(A)} \hat{L}_n(A_j) \frac{N(A_j)}{N(A)}.$$

The quantity $\hat{L}_n(A, k^+)$ is interpreted as the total (normalized) error of a majority vote over the offspring of A living k^+ generations later. It should

be stressed that **both** $\hat{L}_n(A)$ and $\hat{L}_n(A, k^+)$ may be evaluated on the basis of the data points falling in A only (no matter what the rest of the tree looks like), thereby respecting the cellular constraint.

Now, let β be a positive real parameter. With this notation, the stopping rule ① takes the following simple form:

> ① Put $\theta = 0$ if
>
> $$\left|\hat{L}_n(A) - \hat{L}_n(A, k^+)\right| \leq \left(\frac{1}{N(A)+1}\right)^{\beta}.$$

In other words, at each cell, the algorithm compares the actual classification error with the total error of the cell offspring k^+ generations later. This bounded lookahead principle suggested by us is quite well-developed in the artificial intelligence literature—see, for example, Pearl's book (1988) on probabilistic reasoning. If the difference is below some well-chosen threshold, then the cellular classification procedure stops and the node returns a terminal signal. Otherwise, the node outputs 2^d sets of data, and the process continues recursively. The protocol stops once all nodes have returned a terminal signal, and final decisions are taken by majority vote. Thus, for $\mathbf{x}$ falling in a terminal node A, the rule is as usual

$$g_n(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i=1}^n \mathbf{1}_{[\mathbf{X}_i \in A, Y_i = 1]} > \sum_{i=1}^n \mathbf{1}_{[\mathbf{X}_i \in A, Y_i = 0]} \\ 0 & \text{otherwise.} \end{cases}$$

In the next section, we prove the following theorem:

Theorem 4.1 *Let g_n be the cellular tree classifier defined above, with $1 - d\alpha - 2\beta > 0$. Assume that the marginal distributions of $\mathbf{X}$ are nonatomic. Then the classification rule g_n is consistent:*

$$\lim_{n \to \infty} \mathbb{E}L(g_n) = L^\star \quad \text{as } n \to \infty.$$

From a technical point of view, this theorem poses a challenge, as there are no conditions on the distribution, and the rectangular cells do in general not shrink to zero. In fact, it is easy to find distributions of $\mathbf{X}$ for which the maximal cell diameter does not tend to zero in probability, even if all is restricted to the unit cube. For distributions with infinite support, there are always cells of infinite diameter. This observation implies that classical consistency proofs, that often use differentiation of measure arguments or rely on asymptotic justifications related to Lebesgue's density theorem, cannot be applied. The proof uses global arguments instead.

For partitions that do not depend upon the Y-values in the data, consistency can be shown by relatively simple means, following for example the arguments given in Devroye et al. (1996). However, our partition and tree depend upon the Y-values in the data. Within the constraints imposed by the cellular model, we believe that this is the first (and only) proof of universal consistency of a Y-dependent cellular tree classifier. On the other hand, we have proposed a model that is a priori too simple to be competitive. There are choices of parameters to be made, and there is absolutely no minimax theory of lower bounds for the rate with which cellular tree classifiers can approach the Bayes error. On the practical side, besides the question of how to efficiently implement the model, it is also clear that the performance of the cellular estimate will be conditional on a good tuning of both parameters α and β. As a first step, a good route to follow is to attack the rate of convergence problem—we expect dependence on the smoothness of $(\mathbf{X}, Y)$—and deduce from this analysis the best parameter choices. In any case, the work ahead is enormous and the road arduous.

5 Proof of Theorem 4.1

5.1 Notation and preliminary results

We start with some notation (see Figure 4). For each level $k \geq 0$, we denote by $\mathcal{P}_k$ the partition represented by the leaves of the underlying full 2^d-ary median-type tree. This partition has 2^{dk} cells and its construction depends on the $\mathbf{X}_i$'s only. The labels Y_i's do not play a role in the building of $\mathcal{P}_k$, though they are involved in making the decision whether to cut a cell or not.

For each $A_j \in \mathcal{P}_k$, we let $N(A_j)$ be the number of $\mathbf{X}_i$'s falling in A_j and note that $\sum_{j=1}^{2^{dk}} N(A_j) \leq n$, with a strict inequality as soon as $k > 0$ (see Remark 4.2). For each level k, $A_k(\mathbf{X})$ denotes the cell of the partition $\mathcal{P}_k$ into which $\mathbf{X}$ falls, and $N(A_k(\mathbf{X}))$ the number of data points falling in this set.

We let μ be the distribution of $\mathbf{X}$ and η the regression function of Y on $\mathbf{X}$. More precisely, for any Borel-measurable set $A \subset \mathbb{R}^d$,

$$\mu(A) = \mathbb{P}\{\mathbf{X} \in A\}$$

and, for any $\mathbf{x} \in \mathbb{R}^d$,

$$\eta(\mathbf{x}) = \mathbb{P}\{Y = 1 | \mathbf{X} = \mathbf{x}\} = \mathbb{E}[Y | \mathbf{X} = \mathbf{x}].$$

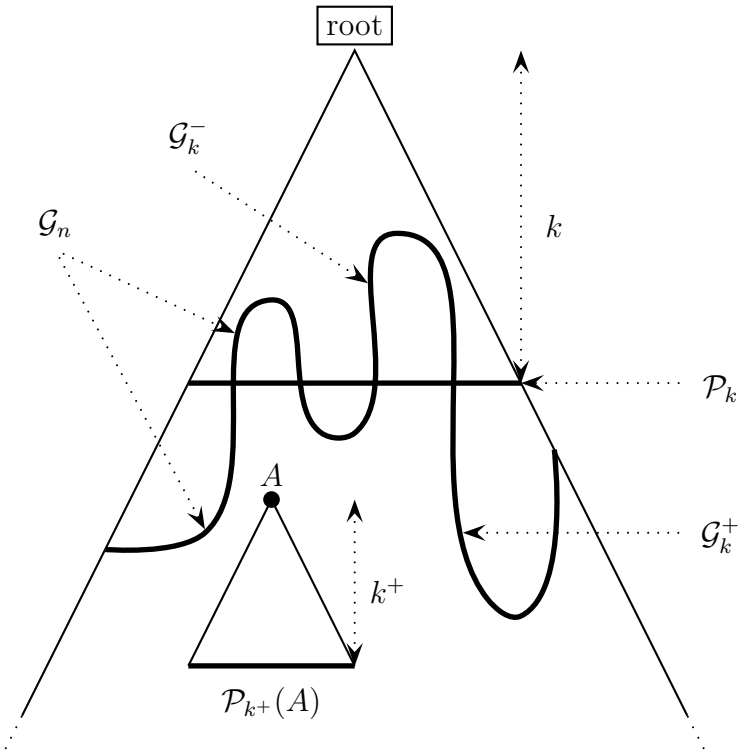

Figure 4: Some key notation.

It is known that the Bayes error is

$$L^\star = \int_{\mathbb{R}^d} \min\left(\eta(\mathbf{z}), 1-\eta(\mathbf{z})\right) \mu(\mathrm{d}\mathbf{z}).$$

Let us recall that, for any cell A,

$$\hat{L}_n(A) = \frac{1}{N(A)} \min\left(\sum_{i=1}^n \mathbf{1}_{[\mathbf{X}_i \in A, Y_i = 1]}, \sum_{i=1}^n \mathbf{1}_{[\mathbf{X}_i \in A, Y_i = 0]}\right).$$

Also, for every $k \geq 0$,

$$\hat{L}_n(A,k) = \sum_{A_j \in \mathcal{P}_k(A)} \hat{L}_n(A_j) \frac{N(A_j)}{N(A)},$$

where $\mathcal{P}_k(A)$ is the full 2^d-ary median-type tree rooted at A of height k. At the population level, we set

$$L^\star(A) = \frac{1}{\mu(A)} \min\left(\int_A \eta(\mathbf{z})\mu(\mathrm{d}\mathbf{z}), \int_A \left(1-\eta(\mathbf{z})\right) \mu(\mathrm{d}\mathbf{z})\right)$$

and

$$L^\star(A,k) = \sum_{A_j \in \mathcal{P}_k(A)} L^\star(A_j) \frac{\mu(A_j)}{\mu(A)}.$$

For all $k \geq 0$, we shall also need the quantity

$$L_k^\star = \mathbb{E}\left[L^\star\left(A_k(\mathbf{X})\right)\right].$$

Note that whenever $A = A(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is a random cell, we take the liberty to abbreviate $\int_A \mathrm{d}\mu$ by $\mu(A)$ throughout the manuscript, since this should cause no confusion. We write for instance

$$L_k^\star = \mathbb{E}\big[\mathbb{E}\left[L^\star\left(A_k(\mathbf{X})\right) \mid \mathbf{X}_1, \ldots, \mathbf{X}_n\right]\big] = \mathbb{E}\left[\sum_{A \in \mathcal{P}_k} L^\star(A)\mu(A)\right]$$

instead of

$$L_k^\star = \mathbb{E}\left[\sum_{A \in \mathcal{P}_k} L^\star(A) \int_A \mathrm{d}\mu\right].$$

Our proof starts with some easy but important facts.

Fact 5.1

(*i*) *For all levels* $k' \geq k \geq 0$,

$$L^\star \leq L_{k'}^\star \leq L_k^\star.$$

(*ii*) *For each cell* A *and each level* $k \geq 0$,

$$\hat{L}_n(A,k) \leq \hat{L}_n(A) + \frac{2^{dk}}{N(A)} \mathbf{1}_{[N(A)>0]}.$$

(*iii*) *For each cell* A *and all levels* $k' \geq k \geq 0$,

$$\hat{L}_n(A,k') \leq \hat{L}_n(A,k) + \frac{2^{dk'}}{N(A)} \mathbf{1}_{[N(A)>0]}.$$

(*iv*) *For each cell* A *and all levels* $k, k' \geq 0$,

$$\mathbb{E}\left[L^\star(A_k(\mathbf{X}), k')\right] = L_{k+k'}^\star.$$

In particular, for $k'' \geq k' \geq 0$,

$$L^\star \leq \mathbb{E}\left[L^\star(A_k(\mathbf{X}), k'')\right] \leq \mathbb{E}\left[L^\star(A_k(\mathbf{X}), k')\right].$$

Proof Proof of statement (i) is based on the nesting of the partitions. To establish (ii), observe that, by definition,

$$\hat{L}_n(A) = \frac{1}{2} - \frac{1}{2N(A)} \left| N(A) - 2 \sum_{i=1}^{n} \mathbf{1}_{[\mathbf{X}_i \in A, Y_i = 1]} \right|,$$

and

$$\begin{aligned}
&\hat{L}_n(A,k) \\
&\quad = \frac{1}{2N(A)} \sum_{A_j \in \mathcal{P}_k(A)} N(A_j) - \frac{1}{2N(A)} \sum_{A_j \in \mathcal{P}_k(A)} \left| N(A_j) - 2 \sum_{i=1}^{n} \mathbf{1}_{[\mathbf{X}_i \in A_j, Y_i = 1]} \right| \\
&\quad \leq \frac{1}{2} - \frac{1}{2N(A)} \sum_{A_j \in \mathcal{P}_k(A)} \left| N(A_j) - 2 \sum_{i=1}^{n} \mathbf{1}_{[\mathbf{X}_i \in A_j, Y_i = 1]} \right|.
\end{aligned}$$

But, by the triangle inequality and Remark 4.2,

$$\begin{aligned}
&\left| N(A) - 2 \sum_{i=1}^{n} \mathbf{1}_{[\mathbf{X}_i \in A, Y_i = 1]} \right| \\
&\qquad \leq \sum_{A_j \in \mathcal{P}_k(A)} \left| N(A_j) - 2 \sum_{i=1}^{n} \mathbf{1}_{[\mathbf{X}_i \in A_j, Y_i = 1]} \right| + 2^{dk} - 1.
\end{aligned}$$

This proves (ii). Proof of (iii) is similar. To show (iv), just note that

$$\begin{aligned}
\mathbb{E}\left[L^\star(A_k(\mathbf{X}), k')\right] &= \mathbb{E}\left[\sum_{A \in \mathcal{P}_k} \sum_{A_j \in \mathcal{P}_{k'}(A)} L^\star(A_j) \frac{\mu(A_j)}{\mu(A)} \mu(A) \right] \\
&= \mathbb{E}\left[L^\star\left(A_{k+k'}(\mathbf{X})\right)\right] \\
&= L^\star_{k+k'}.
\end{aligned}$$

■

The next two propositions will be decisive in our analysis. Proposition 5.1 asserts that the diameter of $A_k(\mathbf{X})$ tends to 0 in probability, provided k (as a function of n) tends sufficiently slowly to infinity. Proposition 5.2 introduces a particular level $k_n^\star$ which will play a central role in the proof of Theorem 4.1.

Proposition 5.1 *Assume that the marginal distributions of* $\mathbf{X}$ *are nonatomic. Then, if*

$$k \to \infty \quad \textit{and} \quad \frac{k 2^{dk}}{n} \to 0,$$

one has

$$\operatorname{diam}\left(A_k(\mathbf{X})\right) \to 0 \quad \textit{in probability as } n \to \infty.$$

Proof of Proposition 5.1 Median-split trees are analyzed in some detail in Section 20.3 of the monograph by Devroye et al. (1996). Starting on page 323, it is shown that the diameter of a randomly selected cell tends to 0 in probability. The adaptation to our 2^d-ary median-type trees is straightforward. However, a few remarks are in order. Section 20.3 of that book assumes that all marginals are uniform. This can also be the set-up for us, because our rule is invariant under monotone transformations of the axes. Note however that it is crucial that splits are made exactly at data points for this property to be true. Also, the proofs in Section 20.3 of Devroye et al. (1996) assume $d = 2$, but are clearly true for general d. The only condition for the diameter result is that of Theorem 20.2, page 323:

$$k \to \infty \quad \text{and} \quad \frac{k2^{dk}}{n} \to \infty.$$

The second condition is only necessary to make sure that the data medians do not run too far away from the true distributional medians. ■

Proposition 5.2 *Let $\psi(n,k)$ be the function defined for all $n \geq 1$ and $k \geq 0$ by*

$$\psi(n,k) = L_k^\star - L^\star.$$

(*i*) *Let $\{k_n\}_{n\geq 1}$ be a sequence of nonnegative integers such that $k_n \to \infty$ and $k_n 2^{dk_n}/n \to 0$. Then*

$$\psi(n,k_n) \to 0 \quad \textit{as } n \to \infty.$$

(*ii*) *Assume that $\alpha \in (0, 1/d)$ and, for fixed n, set*

$$k_n^\star = \min\left\{\ell \geq 0 : \psi(n,\ell) < \sqrt{\left(\frac{2^{d\ell}}{n}\right)^{1-d\alpha}}\right\}.$$

Then

$$\frac{2^{dk_n^\star}}{n} \to 0 \quad \textit{as } n \to \infty.$$

Proof of Proposition 5.2 At first we note, according to Fact 5.1(*ii*), that for all $n \geq 1$ and $k \geq 0$, $\psi(n,k) \geq 0$. For $\mathbf{x} \in \mathbb{R}^d$, introduce

$$\bar{\eta}_n(\mathbf{x}) = \frac{1}{\mu\left(A_{k_n}(\mathbf{x})\right)} \int_{A_{k_n}(\mathbf{x})} \eta(\mathbf{z})\mu(\mathrm{d}\mathbf{z}).$$

With this notation,

$$\begin{aligned}\psi(n,k) &= \mathbb{E}\left[L^\star\left(A_{k_n}(\mathbf{X})\right)\right] - L^\star \\ &\leq \mathbb{E}\left|\eta(\mathbf{X}) - \bar{\eta}_n(\mathbf{X})\right| + \mathbb{E}\left|(1-\eta(\mathbf{X})) - (1-\bar{\eta}_n(\mathbf{X}))\right|.\end{aligned}$$

Let us prove that the first of the two terms above tends to 0 as n tends to infinity—the second term is handled similarly. To this aim, fix $\varepsilon > 0$ and find a uniformly continuous function η_ε on a bounded set $\mathcal{C}$ and vanishing off $\mathcal{C}$ so that $\mathbb{E}|\eta(\mathbf{X}) - \eta_\varepsilon(\mathbf{X})| < \varepsilon$. Clearly, by the triangle inequality,

$$\begin{aligned}\mathbb{E}\left|\eta(\mathbf{X}) - \bar{\eta}_n(\mathbf{X})\right| &\leq \mathbb{E}\left|\eta(\mathbf{X}) - \eta_\varepsilon(\mathbf{X})\right| \\ &\quad + \mathbb{E}\left|\eta_\varepsilon(\mathbf{X}) - \bar{\eta}_{n,\varepsilon}(\mathbf{X})\right| \\ &\quad + \mathbb{E}\left|\bar{\eta}_{n,\varepsilon}(\mathbf{X}) - \bar{\eta}_n(\mathbf{X})\right| \\ &\stackrel{\text{def}}{=} \mathrm{I} + \mathrm{II} + \mathrm{III},\end{aligned}$$

where

$$\bar{\eta}_{n,\varepsilon}(\mathbf{x}) = \frac{1}{\mu\left(A_{k_n}(\mathbf{x})\right)} \int_{A_{k_n}(\mathbf{x})} \eta_\varepsilon(\mathbf{z}) \mu(\mathrm{d}\mathbf{z}).$$

By choice of η_ε, one has $\mathrm{I} < \varepsilon$. Next, note that

$$\mathrm{II} \leq \mathbb{E}\left[\frac{\displaystyle\int_{A_{k_n}(\mathbf{X})} \left|\eta_\varepsilon(\mathbf{X}) - \eta_\varepsilon(\mathbf{z})\right| \mu(\mathrm{d}\mathbf{z})}{\mu\left(A_{k_n}(\mathbf{X})\right)}\right].$$

As η_ε is uniformly continuous, there exists a number $\delta = \delta(\varepsilon) > 0$ such that if $\mathrm{diam}(A) \leq \delta$, then $|\eta_\varepsilon(\mathbf{x}) - \eta_\varepsilon(\mathbf{z})| < \varepsilon$ for every $\mathbf{x}, \mathbf{z} \in A$. In addition, there is a positive constant M such that $|\eta_\varepsilon(\mathbf{x})| \leq M$ for every $\mathbf{x} \in \mathbb{R}^d$. Thus,

$$\mathrm{II} < \varepsilon + 2M\,\mathbb{P}\left\{\mathrm{diam}\left(A_{k_n}(\mathbf{X})\right) > \delta\right\}.$$

Therefore, $\mathrm{II} < 2\varepsilon$ for all n large enough by Proposition 5.1. Finally, $\mathrm{III} \leq \mathrm{I} < \varepsilon$. Taken together, these steps prove the first statement of the proposition.

Next, suppose assertion (ii) is false and set, to simplify notation, $\delta = 1 - d\alpha > 0$. Then we can find a subsequence $\{k^\star_{n_i}\}_{i\geq 1}$ of $\{k^\star_n\}_{n\geq 1}$ and a positive constant C such that, for all i,

$$\frac{2^{dk^\star_{n_i}}}{n_i} \geq C.$$

Since $n_i \to \infty$, it can be assumed, without loss of generality, that $n_i \geq 2$ and $\log_2(Cn_i) \geq 2d$ for all i. This implies in particular

$$\begin{aligned} k^\star_{n_i} - 1 &\geq \frac{\log_2(Cn_i)}{d} - 1 \\ &\geq \frac{\log_2(Cn_i)}{2d}, \end{aligned} \tag{5.1}$$

and $k^\star_{n_i} \geq 2$ as well.

On the one hand, by the very definition of $k^\star_{n_i}$,

$$\begin{aligned} \psi(n_i, k^\star_{n_i} - 1) &\geq \sqrt{\left(\frac{2^{d(k^\star_{n_i}-1)}}{n_i}\right)^\delta} \\ &\geq \sqrt{\frac{C^\delta}{2^{d\delta}}}. \end{aligned} \tag{5.2}$$

On the other hand, by (5.1) and the monotonicity of $\psi(n_i, .)$ (Fact 5.1(ii)), we may write

$$\psi(n_i, k^\star_{n_i} - 1) \leq \psi\left(n_i, \frac{\log_2(Cn_i)}{2d}\right).$$

But, setting

$$t_{n_i} = \frac{\log_2(Cn_i)}{2d},$$

we have

$$\frac{t_{n_i} 2^{dt_{n_i}}}{n_i} = \frac{\log_2(Cn_i)}{2d}\sqrt{\frac{C}{n_i}}.$$

This quantity goes to 0 as $n_i \to \infty$. Moreover, $t_{n_i} \to \infty$ and thus, according to the first statement of the proposition,

$$\psi(n_i, k^\star_{n_i} - 1) \to 0 \quad \text{as } n_i \to \infty.$$

This contradicts (5.2). ■

5.2 Proof of the theorem

Let $\{k^\star_n\}_{n\geq 1}$ be defined as in Proposition 5.2. We denote by $\mathcal{G}_n$ the leaf regions of the cellular tree, and by $\mathcal{G}^-_{k^\star_n}$ (respectively, $\mathcal{G}^+_{k^\star_n}$) the collection of leaves at level at most (respectively, strictly at least) $k^\star_n$. Finally, for any cell A, we set

$$L_n(A) = \mathbb{P}\{g_n(\mathbf{X}) \neq Y, \mathbf{X} \in A \,|\, \mathcal{D}_n\}.$$

With this notation, we have

$$L^\star \leq \mathbb{E}L(g_n) = \mathbb{E}\left[\sum_{A\in\mathcal{G}_n} L_n(A)\right]$$
$$= \mathbb{E}\left[\sum_{A\in\mathcal{G}^-_{k^\star_n}} L_n(A)\right] + \mathbb{E}\left[\sum_{A\in\mathcal{G}^+_{k^\star_n}} L_n(A)\right].$$

Set

$$\varphi(A) = \left(\frac{1}{N(A)+1}\right)^\beta.$$

Then, clearly,

$$\mathbb{E}\left[\sum_{A\in\mathcal{G}^+_{k^\star_n}} L_n(A)\right] \leq \mathbb{E}\left[\sum_{A\in\mathcal{G}^+_{k^\star_n}} \mu(A)\right]$$
$$\leq \mathbb{E}\left[\sum_{A\in\mathcal{P}_{k^\star_n}} \mathbf{1}_{[|\hat{L}_n(A)-\hat{L}_n(A,k^+)|>\varphi(A)]}\mu(A)\right]$$
$$= \mathbb{P}\left\{\left|\hat{L}_n\left(A_{k^\star_n}(\mathbf{X})\right) - \hat{L}_n(A_{k^\star_n}(\mathbf{X}),k^+)\right| > \varphi\left(A_{k^\star_n}(\mathbf{X})\right)\right\}.$$

In the second inequality, we used the definition of the stopping rule of the cellular tree. Therefore, according to technical Lemma 6.5,

$$\mathbb{E}\left[\sum_{A\in\mathcal{G}^+_{k^\star_n}} L_n(A)\right] \leq \mathrm{O}\left(\sqrt{\left(\frac{2^{dk^\star_n}}{n}\right)^{1-d\alpha-2\beta}}\right).$$

Since $1-d\alpha-2\beta > 0$, this term tends to 0 as $n\to\infty$ by the second statement of Proposition 5.2. Next, introduce the notation

$$N_0(A) = \sum_{i=1}^n \mathbf{1}_{[\mathbf{X}_i\in A, Y_i=0]} \quad \text{and} \quad N_1(A) = \sum_{i=1}^n \mathbf{1}_{[\mathbf{X}_i\in A, Y_i=1]},$$

and observe that

$$\mathbb{E}\left[\sum_{A\in\mathcal{G}^-_{k^\star_n}} L_n(A)\right] = \mathbb{E}\Bigg[\sum_{A\in\mathcal{G}^-_{k^\star_n}} \Bigg\{\mathbf{1}_{[N_0(A)\geq N_1(A)]}\int_A \eta(\mathbf{z})\mu(\mathrm{d}\mathbf{z})$$
$$+ \mathbf{1}_{[N_0(A)<N_1(A)]}\int_A (1-\eta(\mathbf{z}))\,\mu(\mathrm{d}\mathbf{z})\Bigg\}\Bigg].$$

For $\mathbf{x}$ falling in the region covered by $\mathcal{G}_{k_n^\star}^-$, denote by $A_{k_n^\star}^-(\mathbf{x})$ the cell of $\mathcal{G}_{k_n^\star}^-$ containing $\mathbf{x}$, and set

$$N(A_{k_n^\star}^-(\mathbf{x})) = \sum_{i=1}^n \mathbf{1}_{[\mathbf{X}_i \in A_{k_n^\star}^-(\mathbf{x})]}.$$

Letting

$$\hat{\eta}_n(\mathbf{x}) = \frac{1}{N\big(A_{k_n^\star}^-(\mathbf{x})\big)} \sum_{i=1}^n \mathbf{1}_{[\mathbf{X}_i \in A_{k_n^\star}^-(\mathbf{x}), Y_i = 1]},$$

we may write

$$\begin{aligned}
&\mathbb{E}\left[\sum_{A \in \mathcal{G}_{k_n^\star}^-} L_n(A)\right] \\
&\quad \le \mathbb{E}\Bigg[\sum_{A \in \mathcal{G}_{k_n^\star}^-} \hat{L}_n(A)\mu(A) \\
&\quad + \sum_{A \in \mathcal{G}_{k_n^\star}^-} \left\{ \mathbf{1}_{[N_0(A) \ge N_1(A)]} \left(\int_A \eta(\mathbf{z})\mu(\mathrm{d}\mathbf{z}) - \int_A \hat{\eta}_n(\mathbf{z})\mu(\mathrm{d}\mathbf{z}) \right) \right\} \\
&\quad + \sum_{A \in \mathcal{G}_{k_n^\star}^-} \left\{ \mathbf{1}_{[N_0(A) < N_1(A)]} \left(\int_A (1-\eta(\mathbf{z}))\,\mu(\mathrm{d}\mathbf{z}) - \int_A (1-\hat{\eta}_n(\mathbf{z}))\,\mu(\mathrm{d}\mathbf{z}) \right) \right\} \Bigg].
\end{aligned}$$

It follows, evoking Lemma 6.6, that

$$\mathbb{E}\left[\sum_{A \in \mathcal{G}_{k_n^\star}^-} L_n(A)\right] \le \mathbb{E}\left[\sum_{A \in \mathcal{G}_{k_n^\star}^-} \hat{L}_n(A)\mu(A)\right] + \mathrm{O}\left(\sqrt{\frac{2^{dk_n^\star}}{n}}\right).$$

The rightmost term tends to 0 according to the second statement of Proposition 5.2.

Thus, to complete the proof, it remains to establish that

$$\mathbb{E}\left[\sum_{A \in \mathcal{G}_{k_n^\star}^-} \hat{L}_n(A)\mu(A)\right] \to L^\star \quad \text{as } n \to \infty.$$

To this aim, observe that by the very definition of $\mathcal{G}^-_{k_n^\star}$, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{A\in\mathcal{G}^-_{k_n^\star}}\hat{L}_n(A)\mu(A)\right] &\le \mathbb{E}\left[\sum_{A\in\mathcal{G}^-_{k_n^\star}}\left(\hat{L}_n(A,k^+)+\varphi(A)\right)\mu(A)\right]\\ &= \mathbb{E}\left[\sum_{A\in\mathcal{G}^-_{k_n^\star}}\hat{L}_n(A,k^+)\mu(A)\right]+\mathbb{E}\left[\sum_{A\in\mathcal{G}^-_{k_n^\star}}\varphi(A)\mu(A)\right]\\ &\stackrel{\text{def}}{=} \mathrm{I}+\mathrm{II}.\end{aligned}$$

For every cell A of $\mathcal{G}^-_{k_n^\star}$, one has

$$\max\left(\frac{n}{2^{dk_n^\star}}-1,1\right)\le N(A)+1\le\frac{n}{2^{dk_n^\star}}+1. \tag{5.3}$$

Therefore, taking n so large that $n/2^{dk_n^\star}>2$ (this is possible by Proposition 5.2(ii)), we obtain

$$\mathrm{II}\le\left(\frac{n}{2^{dk_n^\star}}-1\right)^{-\beta}\mathbb{E}\left[\sum_{A\in\mathcal{G}^-_{k_n^\star}}\mu(A)\right]\le\left(\frac{n}{2^{dk_n^\star}}-1\right)^{-\beta}.$$

Applying Proposition 5.2(ii) again, we conclude that $\mathrm{II}\to 0$ as $n\to\infty$.

Next, define

$$k_n=\left\lfloor\alpha\log_2\left(\frac{n}{2^{dk_n^\star}}-1\right)\right\rfloor\quad\text{and}\quad k_n'=\left\lfloor\alpha\log_2\left(\frac{n}{2^{dk_n^\star}}+1\right)\right\rfloor.$$

Inequality (5.3) implies that for every $A\in\mathcal{G}^-_{k_n^\star}$ and all n large enough,

$$k_n\le k^+\le k_n'.$$

Thus, by Fact 5.1(iii),

$$\begin{aligned}\mathrm{I}&\le\mathbb{E}\left[\sum_{A\in\mathcal{G}^-_{k_n^\star}}\hat{L}_n\left(A,k_n\right)\mu(A)\right]+\mathbb{E}\left[\sum_{A\in\mathcal{G}^-_{k_n^\star}}\frac{2^{dk_n'}}{N(A)}\mu(A)\right]\\ &=\mathbb{E}\left[\sum_{A\in\mathcal{G}^-_{k_n^\star}}\hat{L}_n\left(A,k_n\right)\mu(A)\right]+\mathrm{O}\left(\left(\frac{2^{dk_n^\star}}{n}\right)^{1-d\alpha}\right).\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{A\in\mathcal{G}^-_{k^\star_n}} \hat{L}_n\left(A,k_n\right)\mu(A)\right] \\
&\quad\le \mathbb{E}\left[\sum_{A\in\mathcal{G}^-_{k^\star_n}} L^\star(A,k_n)\mu(A)\right] + \mathbb{E}\left[\sum_{A\in\mathcal{G}^-_{k^\star_n}} \left|\hat{L}_n(A,k_n) - L^\star(A,k_n)\right|\mu(A)\right] \\
&\quad= \mathbb{E}\left[\sum_{A\in\mathcal{G}^-_{k^\star_n}} L^\star(A,k_n)\mu(A)\right] + \mathrm{O}\left(\sqrt{\left(\frac{2^{dk^\star_n}}{n}\right)^{1-d\alpha}}\right) \\
&\qquad\text{(by Lemma 6.4).}
\end{aligned}$$

Consequently,

$$I \le \mathbb{E}\left[\sum_{A\in\mathcal{G}^-_{k^\star_n}} L^\star(A,k_n)\mu(A)\right] + \mathrm{O}\left(\sqrt{\left(\frac{2^{dk^\star_n}}{n}\right)^{1-d\alpha}}\right),$$

and the rightmost term tends to 0 as $n\to\infty$ by Proposition 5.2(ii). Thus, the proof will be finalized if we show that

$$\mathbb{E}\left[\sum_{A\in\mathcal{G}^-_{k^\star_n}} L^\star(A,k_n)\mu(A)\right] \to L^\star \quad \text{as } n\to\infty.$$

We have

$$\begin{aligned}
\mathbb{E}\left[\sum_{A\in\mathcal{G}^-_{k^\star_n}} L^\star(A,k_n)\mu(A)\right] &= \mathbb{E}\left[\sum_{A\in\mathcal{G}^-_{k^\star_n}}\sum_{A_j\in\mathcal{P}_{k_n}(A)} L^\star(A_j)\frac{\mu(A_j)}{\mu(A)}\mu(A)\right] \\
&\le \mathbb{E}\left[\sum_{A\in\mathcal{P}_{k_n}} L^\star(A)\mu(A)\right] \\
&= L^\star_{k_n},
\end{aligned}$$

where, in the inequality, we use the fact that the cells in the double sum are at level at least k_n. But, clearly,

$$\frac{k_n 2^{dk_n}}{n} \le \frac{\alpha\log_2 n}{n^{1-d\alpha}},$$

and consequently, since $d\alpha < 1$,

$$\frac{k_n 2^{dk_n}}{n} \to 0 \quad \text{as } n \to \infty.$$

Thus, by Proposition 5.2(i), the term $L^\star_{k_n}$ tends to $L^\star$. This concludes the proof.

6 Some technical results

Throughout this section, we adopt the general notation of the document. In particular, we let α and β be two positive real numbers such that $1-d\alpha-2\beta > 0$. The sequence $\{k^\star_n\}_{n\geq 1}$ is defined as in Proposition 5.2 and we set

$$k^+ = \lfloor \alpha \log_2(N(A)+1) \rfloor . \tag{6.1}$$

We will repeatedly use the fact that, by Proposition 5.2(ii), $2^{dk^\star_n}/n \to 0$ as $n \to \infty$. For any $k \geq 0$, $\mathcal{T}_k$ stands for the full 2^d-ary median-type tree with k levels of nodes, whose leaves represent $\mathcal{P}_k$.

Recall that $\mathbf{X}$ has probability measure μ on $\mathbb{R}^d$ and that its marginals are assumed to be nonatomic. The first important result that is needed here is the following one:

Proposition 6.1 *Let $\{k_n\}_{n\geq 1}$ be a sequence of nonnegative integers such that $2^{dk_n}/n \to 0$. Then*

$$\mathbb{E}\left[\sum_{A\in\mathcal{P}_{k_n}} \left|\frac{N(A)}{n} - \mu(A)\right|\right] = \mathrm{O}\left(\sqrt{\frac{2^{dk_n}}{n}}\right).$$

Proof of Proposition 6.1 In the sequel, we let n be large enough to ensure that $n/2^{dk_n} > 2$, so that we do not have to worry about empty cells.

To prove the lemma, recall the construction of $\mathcal{T}_{k_n}$. At the root, which represents $\mathbb{R}^d$, we order the points by the first component. We define the pivot as the r-th smallest point, where $r = \lfloor (n+1)/2 \rfloor$, and cut perpendicularly to the first component at the pivot. Let the pivot's first component have value $x^\star$. Define

$$A = \left\{\mathbf{x} \in \mathbb{R}^d : \mathbf{x} = (x_1, \cdots, x_d), x_1 < x^\star\right\}$$

and

$$B = \left\{\mathbf{x} \in \mathbb{R}^d : \mathbf{x} = (x_1, \cdots, x_d), x_1 > x^\star\right\}.$$

The sample points that fall in A, conditionally on the pivot, are distributed according to μ restricted to A, and similarly for B. Also, importantly,

$$\mu(A) \stackrel{\mathcal{L}}{=} \text{Beta}(r, n - r + 1)$$

and

$$\mu(B) \stackrel{\mathcal{L}}{=} \text{Beta}(n - r + 1, r),$$

from the theory of order statistics (see, e.g., David and Nagaraja, 2003).

We need to see how large $\mu(A)$, $\mu(B)$, $N(A)$ and $N(B)$ are. To this aim, we distinguish between the cases where n is odd and n is even.

1. n **odd**. Now $r = (n + 1)/2$, $N(A) = r - 1 = (n - 1)/2$, $N(B) = n - r = (n - 1)/2$, and

$$\mu(A) \stackrel{\mathcal{L}}{=} \mu(B) \stackrel{\mathcal{L}}{=} \text{Beta}\left(\frac{n+1}{2}, \frac{n+1}{2}\right).$$

2. n **even**. In this case we have $r = n/2$, $N(A) = (n-2)/2$, $N(B) = n/2$,

$$\mu(A) \stackrel{\mathcal{L}}{=} \text{Beta}\left(\frac{n}{2}, \frac{n+2}{2}\right)$$

and

$$\mu(B) \stackrel{\mathcal{L}}{=} \text{Beta}\left(\frac{n+2}{2}, \frac{n}{2}\right).$$

As $N(A) + N(B) = n - 1$, the pivot is not sent down to the subtrees. Let us have a canonical way of deciding who goes left and right, e.g., A is left and B is right. Next, still at the root, we rotate the coordinate and repeat the median splitting process for the sample points in A and B (both open sets) in direction 2, then in direction 3, and so forth until direction d. We create this way the 2^d children of the root and, repeating this scheme for k_n levels of nodes, we construct the 2^d-ary tree up to distance k_n from the root. It has exactly 2^{dk_n} leaves.

On any path of length k_n to one of the 2^{dk_n} leaves, we have a deterministic sequence of cardinalities

$$n_0 = n(\text{root}), n_1, n_2, \ldots, n_{k_n}.$$

We have already seen that, for all $i = 0, \ldots, k_n$,

$$\frac{n}{2^{di}} - 2 \leq n_i \leq \frac{n}{2^{di}}.$$

Now, consider a fixed path to a fixed leaf, $(n_0, n_1, \ldots, n_{k_n})$. Then, conditionally on the pivots, the set of $\mathbb{R}^d$ that corresponds to that leaf, i.e., a hyperrectangle of $\mathbb{R}^d$, has μ-measure distributed as

$$\begin{aligned}\mathrm{Beta}(n_1+1, n_0-n_1) \times \cdots \times \mathrm{Beta}(n_{k_n}+1, n_{k_n-1}-n_{k_n}) &\stackrel{\text{def}}{=} Z_1 \times \cdots \times Z_{k_n} \\ &\stackrel{\text{def}}{=} Z.\end{aligned}$$

Observe that

$$\mathbb{E}Z = \prod_{i=1}^{k_n} \mathbb{E}Z_i = \prod_{i=1}^{k_n} \frac{n_i+1}{n_{i-1}+1} = \frac{n_{k_n}+1}{n+1}.$$

Also,

$$\mathbb{E}Z^2 = \prod_{i=1}^{k_n} \mathbb{E}Z_i^2 = \prod_{i=1}^{k_n} \frac{(n_i+1)(n_i+2)}{(n_{i-1}+1)(n_{i-1}+2)} = \frac{(n_{k_n}+1)(n_{k_n}+2)}{(n+1)(n+2)}.$$

The objective is to bound

$$\begin{aligned}\mathbb{E}\left|\frac{n_{k_n}}{n} - Z\right| &\le \sqrt{\mathbb{E}\left|Z - \frac{n_{k_n}}{n}\right|^2} \\ &= \sqrt{\mathbb{E}\left|Z - \mathbb{E}Z\right|^2 + \left|\frac{n_{k_n}}{n} - \mathbb{E}Z\right|^2} \\ &= \sqrt{\mathbb{V}Z + \left|\frac{n_{k_n}}{n} - \mathbb{E}Z\right|^2} \\ &= \sqrt{\mathbb{V}Z + \left|\frac{n_{k_n}}{n} - \frac{n_{k_n}+1}{n+1}\right|^2},\end{aligned}$$

where the symbol $\mathbb{V}$ stands for the variance. Note

$$\left|\frac{n_{k_n}}{n} - \frac{n_{k_n}+1}{n+1}\right| = \left|\frac{n_{k_n}-n}{n(n+1)}\right| \le \frac{1}{n+1}.$$

Also,

$$\begin{aligned}\mathbb{V}Z &= \left(\frac{n_{k_n}+1}{n+1}\right)\left(\frac{n_{k_n}+2}{n+2} - \frac{n_{k_n}+1}{n+1}\right) \\ &= \left(\frac{n_{k_n}+1}{n+1}\right) \times \frac{n-n_{k_n}}{(n+2)(n+1)} \\ &\le \frac{n_{k_n}+1}{(n+1)(n+2)}.\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}\left|\frac{n_{k_n}}{n} - Z\right| &\le \sqrt{\frac{n_{k_n}+1}{(n+1)(n+2)} + \frac{1}{(n+1)^2}} \\ &\le \frac{1}{n+1}\sqrt{n_{k_n}+2}.\end{aligned}$$

Sum over all 2^{dk_n} sets in the partition $\mathcal{P}_{k_n}$, and call the set cardinalities $n_{k_n}(1), \ldots, n_{k_n}(2^{dk_n})$. Then, denoting by Z_i the "Z" for the i-th set in the partition, we obtain

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{2^{dk_n}}\left|\frac{n_{k_n}(i)}{n} - Z_i\right|\right] &\le \frac{1}{n+1}\sum_{i=1}^{2^{dk_n}}\sqrt{n_{k_n}(i)+2} \\ &\le \frac{1}{n+1}\sqrt{\sum_{i=1}^{2^{dk_n}}1}\sqrt{\sum_{i=1}^{2^{dk_n}}(n_{k_n}(i)+2)} \\ &\qquad \text{(by the Cauchy-Schwarz inequality).}\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{2^{dk_n}}\left|\frac{n_{k_n}(i)}{n} - Z_i\right|\right] &\le \frac{\sqrt{2^{dk_n}}}{n+1} \times \sqrt{n + 2^{dk_n+1}} \\ &\le \frac{\sqrt{2^{dk_n}}}{n+1}\left(\sqrt{n} + \sqrt{2^{dk_n+1}}\right) \\ &\le \sqrt{\frac{2^{dk_n}}{n}} + \frac{2^{dk_n+1}}{n}.\end{aligned}$$

Since $2^{dk_n}/n \to 0$ as $n \to \infty$, this last term is $\mathrm{O}(\sqrt{2^{dk_n}/n})$. ■

Corollary 6.1 *Let $\{k_n\}_{n\ge1}$ be a sequence of nonnegative integers such that $2^{dk_n}/n \to 0$, and let $\mathcal{P}_{k_n}^-$ be the partition of $\mathbb{R}^d$ corresponding to the leaves of any subtree of $\mathcal{T}_{k_n}$ rooted at $\mathbb{R}^d$. Then*

$$\mathbb{E}\left[\sum_{A\in\mathcal{P}_{k_n}^-}\left|\frac{N(A)}{n} - \mu(A)\right|\right] = \mathrm{O}\left(\sqrt{\frac{2^{dk_n}}{n}}\right).$$

Proof of Corollary 6.1 The proof is similar to the proof of Proposition 6.1—just note that $\mathcal{P}_{k_n}^-$ has at most 2^{dk_n} cells. ■

Proposition 6.2 *Let $\{k_n\}_{n\geq 1}$ be a sequence of nonnegative integers such that $2^{dk_n}/n \to 0$. Then*

$$\mathbb{E}\left|\frac{1}{N\left(A_{k_n}(\mathbf{X})\right)}\sum_{i=1}^{n}\mathbf{1}_{[\mathbf{X}_i\in A_{k_n}(\mathbf{X}),Y_i=1]}-\frac{1}{\mu\left(A_{k_n}(\mathbf{X})\right)}\int_{A_{k_n}(\mathbf{X})}\eta(\mathbf{z})\mu(\mathrm{d}\mathbf{z})\right| = \mathrm{O}\left(\sqrt{\frac{2^{dk_n}}{n}}\right)$$

and, similarly,

$$\mathbb{E}\left|\frac{1}{N\left(A_{k_n}(\mathbf{X})\right)}\sum_{i=1}^{n}\mathbf{1}_{[\mathbf{X}_i\in A_{k_n}(\mathbf{X}),Y_i=0]}-\frac{1}{\mu\left(A_{k_n}(\mathbf{X})\right)}\int_{A_{k_n}(\mathbf{X})}\left(1-\eta(\mathbf{z})\right)\mu(\mathrm{d}\mathbf{z})\right| = \mathrm{O}\left(\sqrt{\frac{2^{dk_n}}{n}}\right).$$

Proof of Proposition 6.2 We only prove the first statement. Since $n/2^{dk_n} \to \infty$ as $n \to \infty$, we can always choose n large enough so that no cell of $\mathcal{P}_{k_n}$ is empty. A quick check of $\mathcal{T}_{k_n}$ reveals that given the pivots (see Proposition 6.1), the points inside each cell are distributed in an i.i.d. manner according to the restriction of μ to the cell. Moreover, conditionally on $\mathbf{X}$ and the pivots, $N(A_{k_n}(\mathbf{X}))$ has a deterministic, fixed value. Thus, setting

$$\bar{\eta}_n(\mathbf{x}) = \frac{1}{\mu\left(A_{k_n}(\mathbf{x})\right)}\int_{A_{k_n}(\mathbf{x})}\eta(\mathbf{z})\mu(\mathrm{d}\mathbf{z}),$$

we obtain, conditionally on $\mathbf{X}$ and the pivots,

$$\begin{aligned}
&\mathbb{E}\left|\frac{1}{N\left(A_{k_n}(\mathbf{X})\right)}\sum_{i=1}^{n}\mathbf{1}_{[\mathbf{X}_i\in A_{k_n}(\mathbf{X}),Y_i=1]}-\frac{1}{\mu\left(A_{k_n}(\mathbf{X})\right)}\int_{A_{k_n}(\mathbf{X})}\eta(\mathbf{z})\mu(\mathrm{d}\mathbf{z})\right| \\
&\leq \sqrt{\frac{\bar{\eta}_n(\mathbf{X})\left(1-\bar{\eta}_n(\mathbf{X})\right)}{N\left(A_{k_n}(\mathbf{X})\right)}} \\
&\leq \frac{1}{2}\sqrt{\frac{1}{N\left(A_{k_n}(\mathbf{X})\right)}} \\
&\leq \frac{1}{2}\sqrt{\frac{1}{\frac{n}{2^{dk_n}}-2}}.
\end{aligned}$$

The result follows from the condition $2^{dk_n}/n \to 0$. ■

Corollary 6.2 *Let $\{k_n\}_{n\geq 1}$ be a sequence of nonnegative integers such that $2^{dk_n}/n \to 0$, and let $\mathcal{P}_{k_n}^-$ be the partition of $\mathbb{R}^d$ corresponding to the leaves of any subtree of $\mathcal{T}_{k_n}$ rooted at $\mathbb{R}^d$. For each $\mathbf{x} \in \mathbb{R}^d$, denote by $A_{k_n}^-(\mathbf{x})$ the cell of $\mathcal{P}_{k_n}^-$ containing $\mathbf{x}$. Then*

$$\mathbb{E}\left|\frac{1}{N\left(A_{k_n}^-(\mathbf{X})\right)}\sum_{i=1}^{n}\mathbf{1}_{[\mathbf{X}_i\in A_{k_n}^-(\mathbf{X}),Y_i=1]}-\frac{1}{\mu\left(A_{k_n}^-(\mathbf{X})\right)}\int_{A_{k_n}^-(\mathbf{X})}\eta(\mathbf{z})\mu(\mathrm{d}\mathbf{z})\right| = \mathrm{O}\left(\sqrt{\frac{2^{dk_n}}{n}}\right)$$

and, similarly,

$$\mathbb{E}\left|\frac{1}{N\left(A_{k_n}^-(\mathbf{X})\right)}\sum_{i=1}^{n}\mathbf{1}_{[\mathbf{X}_i\in A_{k_n}^-(\mathbf{X}),Y_i=0]}-\frac{1}{\mu\left(A_{k_n}^-(\mathbf{X})\right)}\int_{A_{k_n}^-(\mathbf{X})}(1-\eta(\mathbf{z}))\,\mu(\mathrm{d}\mathbf{z})\right| = \mathrm{O}\left(\sqrt{\frac{2^{dk_n}}{n}}\right).$$

Proof of Corollary 6.2 The proof is similar to that of Proposition 6.2—just note that

$$N\left(A_{k_n}^-(\mathbf{X})\right) \geq \frac{n}{2^{dk_n}}$$

for all n large enough. ■

Lemma 6.1 *Let $\{k_n\}_{n\geq 1}$ be a sequence of nonnegative integers such that $2^{dk_n}/n \to 0$. Then*

$$\mathbb{E}\left|\hat{L}_n\left(A_{k_n}(\mathbf{X})\right)-L^\star\left(A_{k_n}(\mathbf{X})\right)\right| = \mathrm{O}\left(\sqrt{\frac{2^{dk_n}}{n}}\right).$$

Proof of Lemma 6.1 Using the definition of $\hat{L}_n(A_{k_n}(\mathbf{X}))$ and $L^\star(A_{k_n}(\mathbf{X}))$, we may write

$$\begin{aligned}
&\mathbb{E}\left|\hat{L}_n\left(A_{k_n}(\mathbf{X})\right)-L^\star\left(A_{k_n}(\mathbf{X})\right)\right| \\
&\leq \mathbb{E}\left|\frac{1}{N\left(A_{k_n}(\mathbf{X})\right)}\sum_{i=1}^{n}\mathbf{1}_{[\mathbf{X}_i\in A_{k_n}(\mathbf{X}),Y_i=1]}-\frac{1}{\mu\left(A_{k_n}(\mathbf{X})\right)}\int_{A_{k_n}(\mathbf{X})}\eta(\mathbf{z})\mu(\mathrm{d}\mathbf{z})\right| \\
&\quad+\mathbb{E}\left|\frac{1}{N\left(A_{k_n}(\mathbf{X})\right)}\sum_{i=1}^{n}\mathbf{1}_{[\mathbf{X}_i\in A_{k_n}(\mathbf{X}),Y_i=0]}\right. \\
&\qquad\left.-\frac{1}{\mu\left(A_{k_n}(\mathbf{X})\right)}\int_{A_{k_n}(\mathbf{X})}(1-\eta(\mathbf{z}))\,\mu(\mathrm{d}\mathbf{z})\right|.
\end{aligned}$$

Each term of the sum goes to 0 by Proposition 6.2. ■

Lemma 6.2 *Let $\{k_n\}_{n\geq 1}$ be a sequence of nonnegative integers such that $2^{dk_n}/n \to 0$, and let $\mathcal{P}_{k_n}^-$ be the partition of $\mathbb{R}^d$ corresponding to the leaves of any subtree of $\mathcal{T}_{k_n}$ rooted at $\mathbb{R}^d$. For each $\mathbf{x} \in \mathbb{R}^d$, denote by $A_{k_n}^-(\mathbf{x})$ the cell of $\mathcal{P}_{k_n}^-$ containing $\mathbf{x}$. Then*

$$\mathbb{E}\left|\hat{L}_n\left(A_{k_n}^-(\mathbf{X})\right) - L^\star\left(A_{k_n}^-(\mathbf{X})\right)\right| = \mathrm{O}\left(\sqrt{\frac{2^{dk_n}}{n}}\right).$$

Proof of Lemma 6.2 The proof is similar to that of Lemma 6.1. It uses Corollary 6.2 instead of Proposition 6.2. ■

Lemma 6.3 *Let*

$$k_n = \left\lfloor \alpha \log_2\left(\frac{n}{2^{dk_n^\star}} + 1\right)\right\rfloor.$$

Then

$$\mathbb{E}\left|\hat{L}_n(A_{k_n^\star}(\mathbf{X}), k_n) - L^\star(A_{k_n^\star}(\mathbf{X}), k_n)\right| = \mathrm{O}\left(\sqrt{\left(\frac{2^{dk_n^\star}}{n}\right)^{1-d\alpha}}\right).$$

Proof of Lemma 6.3 We have

$$\begin{aligned}
&\mathbb{E}\left|\hat{L}_n(A_{k_n^\star}(\mathbf{X}), k_n) - L^\star(A_{k_n^\star}(\mathbf{X}), k_n)\right| \\
&\quad= \mathbb{E}\left[\sum_{A\in\mathcal{P}_{k_n^\star}}\sum_{A_j\in\mathcal{P}_{k_n}(A)}\left|\hat{L}_n(A_j)\frac{N(A_j)}{N(A)} - L^\star(A_j)\frac{\mu(A_j)}{\mu(A)}\right|\mu(A)\right] \\
&\quad\leq \mathbb{E}\left[\sum_{A\in\mathcal{P}_{k_n^\star}}\sum_{A_j\in\mathcal{P}_{k_n}(A)}\left|\hat{L}_n(A_j)\frac{N(A_j)}{N(A)} - \hat{L}_n(A_j)\frac{\mu(A_j)}{\mu(A)}\right|\mu(A)\right] \\
&\qquad+ \mathbb{E}\left[\sum_{A\in\mathcal{P}_{k_n^\star}}\sum_{A_j\in\mathcal{P}_{k_n}(A)}\left|\hat{L}_n(A_j) - L^\star(A_j)\right|\mu(A_j)\right] \\
&\quad\stackrel{\text{def}}{=} \mathrm{I} + \mathrm{II}.
\end{aligned}$$

Clearly,

$$\mathrm{II} = \mathbb{E}\left|\hat{L}_n\left(A_{k_n^\star+k_n}(\mathbf{X})\right) - L^\star\left(A_{k_n^\star+k_n}(\mathbf{X})\right)\right|$$

whence, according to Lemma 6.1,

$$\mathrm{II} = \mathrm{O}\left(\sqrt{\left(\frac{2^{dk_n^\star}}{n}\right)^{1-d\alpha}}\right).$$

On the other hand, since $\hat{L}_n(A_j) \leq 1$,

$$\begin{aligned}
\mathrm{I} &\leq \mathbb{E}\left[\sum_{A\in\mathcal{P}_{k_n^\star}}\sum_{A_j\in\mathcal{P}_{k_n}(A)}\left|\frac{N(A_j)}{N(A)}-\frac{\mu(A_j)}{\mu(A)}\right|\mu(A)\right] \\
&\leq \mathbb{E}\left[\sum_{A\in\mathcal{P}_{k_n^\star}}\sum_{A_j\in\mathcal{P}_{k_n}(A)}\left|\frac{N(A_j)}{N(A)}\mu(A)-\frac{N(A_j)}{n}\right|\right] \\
&\quad + \mathbb{E}\left[\sum_{A\in\mathcal{P}_{k_n^\star}}\sum_{A_j\in\mathcal{P}_{k_n}(A)}\left|\frac{N(A_j)}{n}-\mu(A_j)\right|\right].
\end{aligned}$$

The inequality

$$\sum_{A_j\in\mathcal{P}_{k_n}(A)} N(A_j) \leq N(A)$$

leads to

$$\begin{aligned}
\mathrm{I} &\leq \mathbb{E}\left[\sum_{A\in\mathcal{P}_{k_n^\star}}\left|\mu(A)-\frac{N(A)}{n}\right|\right] + \mathbb{E}\left[\sum_{A\in\mathcal{P}_{k_n^\star}}\sum_{A_j\in\mathcal{P}_{k_n}(A)}\left|\frac{N(A_j)}{n}-\mu(A_j)\right|\right] \\
&= \mathbb{E}\left[\sum_{A\in\mathcal{P}_{k_n^\star}}\left|\mu(A)-\frac{N(A)}{n}\right|\right] + \mathbb{E}\left[\sum_{A\in\mathcal{P}_{k_n^\star+k_n}}\left|\frac{N(A)}{n}-\mu(A)\right|\right].
\end{aligned}$$

Thus, by Proposition 6.1,

$$\mathrm{I} = \mathrm{O}\left(\sqrt{\left(\frac{2^{dk_n^\star}}{n}\right)^{1-d\alpha}}\right).$$

Collecting bounds, we obtain

$$\mathrm{I}+\mathrm{II} = \mathrm{O}\left(\sqrt{\left(\frac{2^{dk_n^\star}}{n}\right)^{1-d\alpha}}\right).$$

■

Lemma 6.4 *Let $\mathcal{G}_{k_n^\star}^-$ be the collection of cells of $\mathcal{G}_n$ at level at most $k_n^\star$, and let*

$$k_n = \left\lfloor \alpha \log_2\left(\max\left(\frac{n}{2^{dk_n^\star}}-1\right),1\right)\right\rfloor.$$

Then

$$\mathbb{E}\left[\sum_{A\in\mathcal{G}^-_{k^\star_n}}\left|\hat{L}_n(A,k_n)-L^\star(A,k_n)\right|\mu(A)\right]=\mathrm{O}\left(\sqrt{\left(\frac{2^{dk^\star_n}}{n}\right)^{1-d\alpha}}\right).$$

Proof of Lemma 6.4 Denote by $\bar{\mathcal{G}}^-_{k^\star_n}$ the cells of $\mathcal{P}_{k^\star_n}$ such that the path from the root to the cell does not cross $\mathcal{G}^-_{k^\star_n}$. By construction, the subset collection

$$\mathcal{P}^-_{k^\star_n}=\mathcal{G}^-_{k^\star_n}\cup\bar{\mathcal{G}}^-_{k^\star_n}$$

is a partition of $\mathbb{R}^d$ represented by a subtree of $\mathcal{T}_{k^\star_n}$ rooted at $\mathbb{R}^d$. Moreover, clearly,

$$\begin{aligned}&\mathbb{E}\left[\sum_{A\in\mathcal{G}^-_{k^\star_n}}\left|\hat{L}_n(A,k_n)-L^\star(A,k_n)\right|\mu(A)\right]\\&\quad\leq\mathbb{E}\left[\sum_{A\in\mathcal{P}^-_{k^\star_n}}\left|\hat{L}_n(A,k_n)-L^\star(A,k_n)\right|\mu(A)\right].\end{aligned}$$

Thus, denoting by $A^-_{k^\star_n}(\mathbf{x})$ the cell of $\mathcal{P}^-_{k^\star_n}$ containing $\mathbf{x}$, we are led to

$$\begin{aligned}&\mathbb{E}\left[\sum_{A\in\mathcal{G}^-_{k^\star_n}}\left|\hat{L}_n(A,k_n)-L^\star(A,k_n)\right|\mu(A)\right]\\&\quad\leq\mathbb{E}\left|\hat{L}_n(A^-_{k^\star_n}(\mathbf{X}),k_n)-L^\star(A^-_{k^\star_n}(\mathbf{X}),k_n)\right|.\end{aligned}$$

The end of the proof is similar to the proof of Lemma 6.3. Replace $\mathcal{P}_{k^\star_n}$ by $\mathcal{P}^-_{k^\star_n}$ and invoke Corollary 6.1 (instead of Proposition 6.1) and Lemma 6.2 (instead of Lemma 6.1). ■

Proposition 6.3 *Let k^+ be defined as in (6.1). Then*

$$\mathbb{E}\left|\hat{L}_n\left(A_{k^\star_n}(\mathbf{X})\right)-\hat{L}_n(A_{k^\star_n}(\mathbf{X}),k^+)\right|\leq\psi(n,k^\star_n)+\mathrm{O}\left(\sqrt{\left(\frac{2^{dk^\star_n}}{n}\right)^{1-d\alpha}}\right),$$

where

$$\psi(n,k)=L^\star_k-L^\star.$$

Proof of Proposition 6.3 For every cell A of $\mathcal{P}_{k_n^\star}^-$, one has

$$\max\left(\frac{n}{2^{dk_n^\star}} - 1, 1\right) \leq N(A) + 1 \leq \frac{n}{2^{dk_n^\star}} + 1. \tag{6.2}$$

Define

$$k_n' = \left\lfloor \alpha \log_2 \left(\frac{n}{2^{dk_n^\star}} - 1 \right) \right\rfloor \quad \text{and} \quad k_n = \left\lfloor \alpha \log_2 \left(\frac{n}{2^{dk_n^\star}} + 1 \right) \right\rfloor,$$

and note that, by inequalities (6.2), for all n large enough,

$$k_n' \leq k^+ \leq k_n.$$

Thus, by the triangle inequality and Fact 5.1(ii), we may write

$$\begin{aligned}
&\mathbb{E}\left|\hat{L}_n\left(A_{k_n^\star}(\mathbf{X})\right) - \hat{L}_n(A_{k_n^\star}(\mathbf{X}), k^+)\right| \\
&\quad \leq \mathbb{E}\left[\hat{L}_n\left(A_{k_n^\star}(\mathbf{X})\right) - \hat{L}_n(A_{k_n^\star}(\mathbf{X}), k^+) + \frac{2^{dk^+}}{N\left(A(\mathbf{X})\right)}\mathbf{1}_{[N(A(\mathbf{X}))>0]}\right] \\
&\qquad + \mathbb{E}\left[\frac{2^{dk^+}}{N\left(A(\mathbf{X})\right)}\mathbf{1}_{[N(A(\mathbf{X}))>0]}\right] \\
&\quad = \mathbb{E}\left[\hat{L}_n\left(A_{k_n^\star}(\mathbf{X})\right) - \hat{L}_n(A_{k_n^\star}(\mathbf{X}), k^+)\right] + \mathrm{O}\left(\left(\frac{2^{dk_n^\star}}{n}\right)^{1-d\alpha}\right).
\end{aligned}$$

Consequently, by Fact 5.1(iii),

$$\begin{aligned}
&\mathbb{E}\left|\hat{L}_n\left(A_{k_n^\star}(\mathbf{X})\right) - \hat{L}_n(A_{k_n^\star}(\mathbf{X}), k^+)\right| \\
&\quad \leq \mathbb{E}\left[\hat{L}_n\left(A_{k_n^\star}(\mathbf{X})\right) - \hat{L}_n(A_{k_n^\star}(\mathbf{X}), k_n)\right] + \mathrm{O}\left(\left(\frac{2^{dk_n^\star}}{n}\right)^{1-d\alpha}\right).
\end{aligned} \tag{6.3}$$

With respect to the first term on the right-hand side, we have

$$\begin{aligned}
&\mathbb{E}\left[\hat{L}_n\left(A_{k_n^\star}(\mathbf{X})\right) - \hat{L}_n(A_{k_n^\star}(\mathbf{X}), k_n)\right] \\
&\quad \leq \mathbb{E}\left|\hat{L}_n\left(A_{k_n^\star}(\mathbf{X})\right) - L^\star\left(A_{k_n^\star}(\mathbf{X})\right)\right| \\
&\qquad + L_{k_n^\star}^\star - L^\star \\
&\qquad + L^\star - \mathbb{E}\left[L^\star(A_{k_n^\star}(\mathbf{X}), k_n)\right] \\
&\qquad + \mathbb{E}\left|L^\star(A_{k_n^\star}(\mathbf{X}), k_n) - \hat{L}_n(A_{k_n^\star}(\mathbf{X}), k_n)\right|.
\end{aligned}$$

According to Lemma 6.1, the first of the four terms above is $\mathrm{O}(\sqrt{2^{dk_n^\star}/n})$, whereas the third one is nonpositive by Fact 5.1(iv). Consequently,

$$\mathbb{E}\left[\hat{L}_n\left(A_{k_n^\star}(\mathbf{X})\right) - \hat{L}_n\left(A_{k_n^\star}(\mathbf{X}), k_n\right)\right] \leq \psi(n, k_n^\star) + \mathrm{O}\left(\sqrt{\frac{2^{dk_n^\star}}{n}}\right) + \mathbb{E}\left|L^\star(A_{k_n^\star}(\mathbf{X}), k_n) - \hat{L}_n(A_{k_n^\star}(\mathbf{X}), k_n)\right|.$$

Evoking finally Lemma 6.3, we see that

$$\mathbb{E}\left[\hat{L}_n\left(A_{k_n^\star}(\mathbf{X})\right) - \hat{L}_n(A_{k_n^\star}(\mathbf{X}), k_n)\right] \leq \psi(n, k_n^\star) + \mathrm{O}\left(\sqrt{\left(\frac{2^{dk_n^\star}}{n}\right)^{1-d\alpha}}\right).$$

Combining this result with (6.3) leads to the desired statement. ■

Lemma 6.5 *Let k^+ be defined as in (6.1). Then*

$$\mathbb{P}\left\{\left|\hat{L}_n\left(A_{k_n^\star}(\mathbf{X})\right) - \hat{L}_n(A_{k_n^\star}(\mathbf{X}), k^+)\right| > \left(\frac{1}{N\left(A_{k_n^\star}(\mathbf{X})\right)+1}\right)^\beta\right\} = \mathrm{O}\left(\sqrt{\left(\frac{2^{dk_n^\star}}{n}\right)^{1-d\alpha-2\beta}}\right).$$

Proof of Lemma 6.5 Set

$$\varphi(A) = \left(\frac{1}{N(A)+1}\right)^\beta.$$

Since $N(A_{k_n^\star}(\mathbf{x})) \leq n/2^{dk_n^\star}$, one has

$$\mathbb{P}\left\{\left|\hat{L}_n\left(A_{k_n^\star}(\mathbf{X})\right) - \hat{L}_n(A_{k_n^\star}(\mathbf{X}), k^+)\right| > \varphi\left(A_{k_n^\star}(\mathbf{X})\right)\right\} \leq \mathbb{P}\left\{\left|\hat{L}_n\left(A_{k_n^\star}(\mathbf{X})\right) - \hat{L}_n(A_{k_n^\star}(\mathbf{X}), k^+)\right| > \left(\frac{1}{n/2^{dk_n^\star}+1}\right)^\beta\right\}.$$

Therefore, by Markov's inequality,

$$\mathbb{P}\left\{\left|\hat{L}_n\left(A_{k_n^\star}(\mathbf{X})\right) - \hat{L}_n(A_{k_n^\star}(\mathbf{X}), k^+)\right| > \varphi\left(A_{k_n^\star}(\mathbf{X})\right)\right\} \leq (n/2^{dk_n^\star}+1)^\beta \times \mathbb{E}\left|\hat{L}_n\left(A_{k_n^\star}(\mathbf{X})\right) - \hat{L}_n(A_{k_n^\star}(\mathbf{X}), k^+)\right|.$$

Thus, by Proposition 6.3,

$$\mathbb{P}\left\{\left|\hat{L}_n\left(A_{k_n^\star}(\mathbf{X})\right) - \hat{L}_n(A_{k_n^\star}(\mathbf{X}), k^+)\right| > \varphi\left(A_{k_n^\star}(\mathbf{X})\right)\right\}$$
$$\leq (n/2^{dk_n^\star}+1)^\beta \times \left[\psi(n,k_n^\star) + \mathrm{O}\left(\sqrt{\left(\frac{2^{dk_n^\star}}{n}\right)^{1-d\alpha}}\right)\right].$$

But, by definition of $k_n^\star$,

$$\psi(n,k_n^\star) < \sqrt{\left(\frac{2^{dk_n^\star}}{n}\right)^{1-d\alpha}}.$$

It follows, since $n/2^{dk_n^\star} \to \infty$, that

$$\mathbb{P}\left\{\left|\hat{L}_n\left(A_{k_n^\star}(\mathbf{X})\right) - \hat{L}_n(A_{k_n^\star}(\mathbf{X}), k^+)\right| > \varphi(A)\right\} = \mathrm{O}\left(\sqrt{\left(\frac{2^{dk_n^\star}}{n}\right)^{1-d\alpha-2\beta}}\right).$$

■

Lemma 6.6 *Let $\mathcal{G}_{k_n^\star}^-$ be the collection of cells of $\mathcal{G}_n$ at level at most $k_n^\star$. For $\mathbf{x} \in \mathbb{R}^d$, denote by $A_{k_n^\star}^-(\mathbf{x})$ the cell of $\mathcal{G}_{k_n^\star}^-$ containing $\mathbf{x}$, and set $N(A_{k_n^\star}^-(\mathbf{x})) = \sum_{i=1}^n \mathbf{1}_{[\mathbf{X}_i \in A_{k_n^\star}^-(\mathbf{x})]}$. Define*

$$\hat{\eta}_n(\mathbf{x}) = \frac{1}{N\left(A_{k_n^\star}^-(\mathbf{x})\right)} \sum_{i=1}^n \mathbf{1}_{[\mathbf{X}_i \in A_{k_n^\star}^-(\mathbf{x}), Y_i = 1]}.$$

Then

$$\mathbb{E}\left[\sum_{A \in \mathcal{G}_{k_n^\star}^-} \left|\int_A \hat{\eta}_n(\mathbf{z})\mu(\mathrm{d}\mathbf{z}) - \int_A \eta(\mathbf{z})\mu(\mathrm{d}\mathbf{z})\right|\right] = \mathrm{O}\left(\sqrt{\frac{2^{dk_n^\star}}{n}}\right)$$

and, similarly,

$$\mathbb{E}\left[\sum_{A \in \mathcal{G}_{k_n^\star}^-} \left|\int_A \left(1-\hat{\eta}_n(\mathbf{z})\right)\mu(\mathrm{d}\mathbf{z}) - \int_A \left(1-\eta(\mathbf{z})\right)\mu(\mathrm{d}\mathbf{z})\right|\right] = \mathrm{O}\left(\sqrt{\frac{2^{dk_n^\star}}{n}}\right).$$

Proof of Lemma 6.6 We only have to prove the first statement. To this aim, observe that

$$\mathbb{E}\left[\sum_{A\in\mathcal{G}_{k_n^\star}^-}\left|\int_A \hat{\eta}_n(\mathbf{z})\mu(\mathrm{d}\mathbf{z}) - \int_A \eta(\mathbf{z})\mu(\mathrm{d}\mathbf{z})\right|\right]$$
$$= \mathbb{E}\left[\sum_{A\in\mathcal{G}_{k_n^\star}^-}\left|\frac{1}{N(A)}\sum_{i=1}^n \mathbf{1}_{[\mathbf{X}_i\in A, Y_i=1]} - \frac{1}{\mu(A)}\int_A \eta(\mathbf{z})\mu(\mathrm{d}\mathbf{z})\right|\mu(A)\right].$$

Denote by $\bar{\mathcal{G}}_{k_n^\star}^-$ the cells of $\mathcal{P}_{k_n^\star}$ such that the path from the root to the cell does not cross $\mathcal{G}_{k_n^\star}^-$. By construction, the subset collection

$$\mathcal{P}_{k_n^\star}^- = \mathcal{G}_{k_n^\star}^- \cup \bar{\mathcal{G}}_{k_n^\star}^-$$

is a partition of $\mathbb{R}^d$ represented by a subtree of $\mathcal{T}_{k_n^\star}$ rooted at $\mathbb{R}^d$. Now,

$$\mathbb{E}\left[\sum_{A\in\mathcal{G}_{k_n^\star}^-}\left|\int_A \hat{\eta}_n(\mathbf{z})\mu(\mathrm{d}\mathbf{z}) - \int_A \eta(\mathbf{z})\mu(\mathrm{d}\mathbf{z})\right|\right]$$
$$\leq \mathbb{E}\left[\sum_{A\in\mathcal{P}_{k_n^\star}^-}\left|\frac{1}{N(A)}\sum_{i=1}^n \mathbf{1}_{[\mathbf{X}_i\in A, Y_i=1]} - \frac{1}{\mu(A)}\int_A \eta(\mathbf{z})\mu(\mathrm{d}\mathbf{z})\right|\mu(A)\right].$$

But, since $\mathcal{P}_{k_n^\star}^-$ is a partition of $\mathbb{R}^d$, one has

$$\mathbb{E}\left[\sum_{A\in\mathcal{P}_{k_n^\star}^-}\left|\frac{1}{N(A)}\sum_{i=1}^n \mathbf{1}_{[\mathbf{X}_i\in A, Y_i=1]} - \frac{1}{\mu(A)}\int_A \eta(\mathbf{z})\mu(\mathrm{d}\mathbf{z})\right|\mu(A)\right]$$
$$= \mathbb{E}\left[\left|\frac{1}{N\big(A_{k_n^\star}^-(\mathbf{X})\big)}\sum_{i=1}^n \mathbf{1}_{[\mathbf{X}_i\in A_{k_n^\star}^-(\mathbf{X}), Y_i=1]} - \frac{1}{\mu\big(A_{k_n^\star}^-(\mathbf{X})\big)}\int_{A_{k_n^\star}^-(\mathbf{X})} \eta(\mathbf{z})\mu(\mathrm{d}\mathbf{z})\right|\right].$$

This term goes to 0 by Corollary 6.2. ■

Acknowledgments We thank the Editor and an anonymous referee for valuable comments and insightful suggestions.

References

A.C. Anderson and K.S. Fu. Design and development of a linear binary tree classifier for leukocytes. Technical Report TR-EE-79-31, Purdue University, 1979.

P. Argentiero, R. Chin, and P. Beaudet. An automated approach to the design of decision tree classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4:51–57, 1982.

A. Arora, P. Dutta, S. Bapat, V. Kulathumani, H. Zhang, V. Naik, V. Mittal, H. Cao, M. Demirbas, M. Gouda, Y. Choi, T. Herman, S. Kulkarni, U. Arumugam, M. Nesterenko, A. Vora, and M. Miyashita. A line in the sand: A wireless sensor network for target detection, classification, and tracking. *Computer Networks*, 46:605–634, 2004.

L.A. Bartolucci, P.H. Swain, and C. Wu. Selective radiant temperature mapping using a layered classifier. *IEEE Transactions on Geosciences and Electronics*, 14:101–106, 1976.

J.L. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18:509–517, 1975.

G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9:2015–2033, 2008.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Chapman & Hall, New York, 1984.

X. Cheng, J. Xu, J. Pei, and J. Liu. Hierarchical distributed data classification in wireless sensor networks. *Computer Communications*, 33:1404–1413, 2010.

P.A. Chou. Optimal partitioning for classification and regression trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:340–354, 1991.

T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. *Introduction to Algorithms. Third Edition*. The MIT Press, Cambridge, 2009.

T.M. Cover and J.A. Thomas. *Elements of Information Theory. Second Edition*. Wiley, New York, 2006.

H.A. David and H.N. Nagaraja. *Order Statistics. Third Edition.* Wiley, Hoboken, 2003.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer, New York, 1996.

M. Drmota. *Random Trees.* Springer, Vienna, 2009.

H. Edelsbrunner and J. van Leeuwen. Multidimensional data structures and algorithms: A bibliography. Technical Report F104, Technische Universität Graz, 1983.

P. Flajolet and R. Sedgewick. *Analytic Combinatorics.* Cambridge University Press, Cambridge, 2008.

J.H. Friedman. A tree-structured approach to nonparametric multiple regression. In T. Gasser and M. Rosenblatt, editors, *Smoothing Techniques for Curve Estimation*, pages 5–22, Heidelberg, 1979. Lecture Notes in Mathematics #757, Springer.

S.B. Gelfand and E.J. Delp. On tree structured classifiers. In I.K. Sethi and A.K. Jain, editors, *Artificial Neural Networks and Statistical Pattern Recognition, Old and New Connections*, pages 71–88, Amsterdam, 1991. Elsevier Science Publishers.

S.B. Gelfand, C.S. Ravishankar, and E.J. Delp. An iterative growing and pruning algorithm for classification tree design. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:163–174, 1991.

S. Gey and E. Nédélec. Model selection for CART regression trees. *IEEE Transactions on Information Theory*, 51:658–670, 2005.

L. Gordon and R.A. Olshen. Asymptotically efficient solutions to the classification problem. *The Annals of Statistics*, 6:515–533, 1978.

H. Guo and S.B. Gelfand. Classification trees with neural network feature extraction. *IEEE Transactions on Neural Networks*, 3:923–933, 1992.

D.E. Gustafson, S. Gelfand, and S.K. Mitter. A nonparametric multiclass partitioning method for classification. In *Proceedings of the Fifth International Conference on Pattern Recognition*, pages 654–659, 1980.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer, New York, 2002.

C.R.P. Hartmann, P.K. Varshney, K.G. Mehrotra, and C.L. Gerberich. Application of information theory to the construction of efficient decision trees. *IEEE Transactions on Information Theory*, 28:565–577, 1982.

M.I. Jordan. A message from the President: The era of big data. *ISBA Bulletin*, 18:1–3, 2011.

M.W. Kurzynski. The optimal strategy of a tree classifier. *Pattern Recognition*, 16:81–87, 1983.

Y.K. Lin and K.S. Fu. Automatic classification of cervical cells using a binary tree classifier. *Pattern Recognition*, 16:69–80, 1983.

W.Y. Loh and N. Vanichsetakul. Tree-structured classification via generalized discriminant analysis. *Journal of the American Statistical Association*, 83:715–728, 1988.

K. Mehlhorn. *Data Structures and Algorithms 3: Multi-dimensional Searching and Computational Geometry*. Springer, Berlin, 1984.

W.S. Meisel and D.A. Michalopoulos. A partitioning algorithm with application in pattern classification and the optimization of decision trees. *IEEE Transactions on Computers*, 22:93–103, 1973.

J.K. Mui and K.S. Fu. Automated classification of nucleated blood cells using a binary tree classifier. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2:429–443, 1980.

M.H. Overmars and J. van Leeuwen. Dynamic multi-dimensional data structures based on quad- and k-d trees. *Acta Informatica*, 17:265–287, 1982.

Y. Park and J. Sklansky. Automated design of linear tree classifiers. *Pattern Recognition*, 23:1393–1412, 1990.

H.J. Payne and W.S. Meisel. An algorithm for constructing optimal binary decision trees. *IEEE Transactions on Computers*, 26:905–916, 1977.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Mateo, 1988.

S. Qing-Yun and K.S. Fu. A method for the design of binary tree classifiers. *Pattern Recognition*, 16:593–603, 1983.

J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, 1993.

L. Rokach and O. Maimon. *Data Mining with Decision Trees: Theory and Applications*. World Scientific, Singapore, 2008.

H. Samet. The quadtree and related hierarchical data structures. *Computing Surveys*, 16:187–260, 1984.

H. Samet. *The Design and Analysis of Spatial Data Structures*. Addison-Wesley, Reading, 1990.

I.K. Sethi and B. Chatterjee. Efficient decision tree design for discrete variable pattern recognition problems. *Pattern Recognition*, 9:197–206, 1977.

S. Shlien. Multiple binary decision tree classifiers. *Pattern Recognition*, 23: 757–763, 1990.

H.U. Simon. The Vapnik-Chervonenkis dimension of decision trees with bounded rank. *Information Processing Letters*, 39:137–141, 1991.

C.Y. Suen and Q.R. Wang. Large tree classifier with heuristic search and global training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9:91–101, 1987.

P.H. Swain and H. Hauska. The decision tree classifier: Design and potential. *IEEE Transactions on Geosciences and Electronics*, 15:142–147, 1977.

Q.R. Wang and C.Y. Suen. Analysis and design of a decision tree based on entropy reduction and its application to large character set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:406–417, 1984.

K.C. You and K.S. Fu. An approach to the design of a linear binary tree classifier. In *Proceedings of the Symposium of Machine Processing of Remotely Sensed Data*, pages 3A–1–10, West Lafayette, 1976. Purdue University.